Error Rates in a Clinical Data Repository: Lessons from the Transition to Electronic Data Transfer

Journal:	*BMJ Open*
Manuscript ID:	bmjopen-2012-002406
Article Type:	Research
Date Submitted by the Author:	26-Nov-2012
Complete List of Authors:	Hong, Matthew KH; Royal Melbourne Hospital, Department of Surgery Yao, Henry Han-I; Royal Melbourne Hospital, Department of Surgery Pedersen, John; TissuPath Specialist Pathology, Peters, Justin; Royal Melbourne Hospital, Department of Urology Costello, Anthony; Royal Melbourne Hospital, Department of Urology Murphy, Declan; Peter MacCallum Cancer Centre, Urological Service Team Hovens, Chris; Royal Melbourne Hospital, Department of Surgery Corcoran, Niall; Royal Melbourne Hospital, Department of Urology
<b>Primary Subject Heading</b>:	Health informatics
Secondary Subject Heading:	Urology, Oncology, Health informatics
Keywords:	database, prostate cancer, data quality, error sources, clinical informatics

Error Rates in a Clinical Data Repository: Lessons from the Transition to Electronic Data Transfer

Matthew KH Hong
Henry HI Yao
John S Pedersen[*]
Justin S Peters
Anthony J Costello
Declan G Murphy[+]
Christopher M Hovens
Niall M Corcoran

Division of Urology, Department of Surgery, University of Melbourne, Royal Melbourne Hospital and the Australian Prostate Cancer Research Centre Epworth, Victoria, Australia.
* TissuPath Specialist Pathology, Mount Waverley and Monash University Faculty of Medicine, Victoria, Australia
+ Peter MacCallum Cancer Centre, East Melbourne, Australia

Correspondence:
Dr Matthew Hong,
Department of Surgery
University of Melbourne
Royal Melbourne Hospital
Grattan St Parkville
VIC 3050 Australia
T: 613 9342 7703
E: m.k.hong@ausdoctors.net

Keywords: Database, prostate cancer, data quality, error sources, clinical informatics
Word count: 3927 words

Article Summary

Article focus

- Although use of structured electronic databases is widespread, a substantial amount of clinical data used in research predates this.
- There is a paucity of literature on error rates in such clinical datasets used in research.
- We explored the reliability of manually transcribed data across different pathology fields in a prostate cancer database and also measured error rates attributable to the source data.

Key messages

- Whilst overall rate of error for manually entered data can be low, individual fields may be variably prone to error, especially those involving descriptive text or requiring an element of interpretation.
- Computerised systems can be used to check clinical source data for error.
- The use of electronic data feeds retrospectively can replace manually collected data fields in some cases to improve overall accuracy.

Strengths and limitations of this study

- Our study design provides a realistic representation of a small to moderate sized oncology database used for research purposes.
- We checked the integrity of one aspect of our source data.
- Our study was limited by its use of a single spreadsheet from a single series of patients.

- As we only examined pathology fields covered by electronic import, the findings were not representative of the entire dataset.

ABSTRACT

Objective: Data errors are a well-documented part of clinical datasets as is their potential to confound downstream analysis. In this study we explore the reliability of manually transcribed data across different pathology fields in a prostate cancer database and also measure error rates attributable to the source data.

Design: Descriptive study

Setting: Specialist urology service at a single centre in metropolitan Victoria

Participants: Between 2004 and 2011, 1471 patients underwent radical prostatectomy at our institution. In a large proportion of these cases, clinicopathological variables were recorded by manual data-entry. In 2011, we obtained electronic versions of the same printed pathology reports for our cohort. The data were electronically imported in parallel to any existing manual entry record enabling direct comparison between them.

Outcome measures: Error rates of manually entered data compared with electronically imported data across clinicopathological fields.

Results: 421 patients had at least 10 comparable pathology fields between the electronic import and manual records and were selected for study. 320 patients had concordant data between manually entered and electronically populated fields in a median of 12 pathology fields (range 10-13), indicating an outright accuracy in manually entered pathology data in 76% of patients. Across all fields, the error rate was 2.8% whilst individual field error ranges from 0.5-6.4%. Fields in text formats were significantly more error-prone than those with direct measurements or involving numerical figures (p<0.001). 971 cases were available for review of error within the source data, with figures of 0.1%-0.9%.

Conclusion: While the overall rate of error was low in manually entered data, individual pathology fields were variably prone to error. High quality pathology data can be obtained for both prospective and retrospective parts of our data repository and the electronic checking of source pathology data for error is feasible.

BACKGROUND AND SIGNIFICANCE

The majority of clinical research publications are based on the analysis of prospectively collected, clinical databases. In addition, patient centred databases are increasingly important in translational research efforts, as appropriately annotated tissue banks are the foundation for global multi-institutional collaborative efforts in genetic and epigenetic screening of various diseases[1]. Yet despite the stringent quality controls placed on the vast amounts of research data derived from these studies and the acute awareness of the need to control data quality[2 3], the inherent accuracy of original clinical datasets is one area that receives relatively little attention.

Data errors are common in clinical datasets[4-6], with some cancer databases recording error rates as high as almost 27% in some fields [7]. Such errors have the potential to adversely affect data analysis and interpretation, and can lead to erroneous conclusions[8]. Methods to first identify then correct errors in these datasets would be immensely valuable in the setting of the large-scale genomics projects being performed.

Two types of errors are described in the literature: one of omission, and one of erroneous value. Although it is sometimes argued that missing values carry greater impact due to their greater prevalence[9], which may be up to 55% in cancer surgery databases[10], these errors are more easily detected with judicious computer queries and corrected with retrospective data collection. On the contrary, once erroneous values permeate a dataset, their effects can cascade in unpredictable ways. Errors in high impact fields have been shown to adversely affect the interpretation of statistical analyses, even if the errors are at low prevalence[11]. Whilst it is well known that

structured data entry improves the accuracy of manual documentation[12], much of the clinical data of high value to researchers predates any effective informatics solutions aimed at data quality that might exist today. Instead, manual retrospective transcription of data from clinical records into relatively unstructured spreadsheets constitutes the data entry method for many clinical audits that subsequently serve research purposes. These datasets may have even transitioned to more carefully constructed data entry interfaces, as might occur in conditions such as prostate cancer where long follow up times of over ten years are necessary for study of oncological outcomes[13]. In such cases, the provenance of the data collected with earlier means may not be accounted for with subsequent analysis.

Studies involving large cancer datasets rarely report error rates or their management, and it is difficult to assess the impact that these may have on the outcomes reported[14]. Given the considerable effort that generally goes towards the collection of data for a large clinical database, it is unsurprising that surplus resources are usually unavailable to place towards the check of data accuracy. Although larger numbers in databases may be used to counter the problem of errors, the combination of datasets, particularly with different fields would serve only to magnify error rates.

Knowledge of errors in manually collected data could give insight into how these may be accounted for in subsequent analysis. In cancer databases, pathology data are of particular importance as they are relied upon to build cohorts of clinical relevance in research. Often there are multiple fields that give equivalent indications of underlying biology and any one could be analysed to similar effect. For example, both the

percentage involvement of tumour or its measured size may serve as parameters of cancer burden. It is unlikely that different types of data would have equivalent vulnerability to error, and knowledge of the fields or types of fields that might be more error-prone with manual data entry could help researchers judiciously select fields for analysis based on greater accuracy. It might also aid informaticians to focus on error prevention in fields that carry particular importance in clinical and research settings. In addition, it is important to measure the baseline level of error inherent within the pathology reports themselves, the data source, as no degree of accuracy in manual transcription or even automated processes can result in a lower error rate without amendment of the original report.

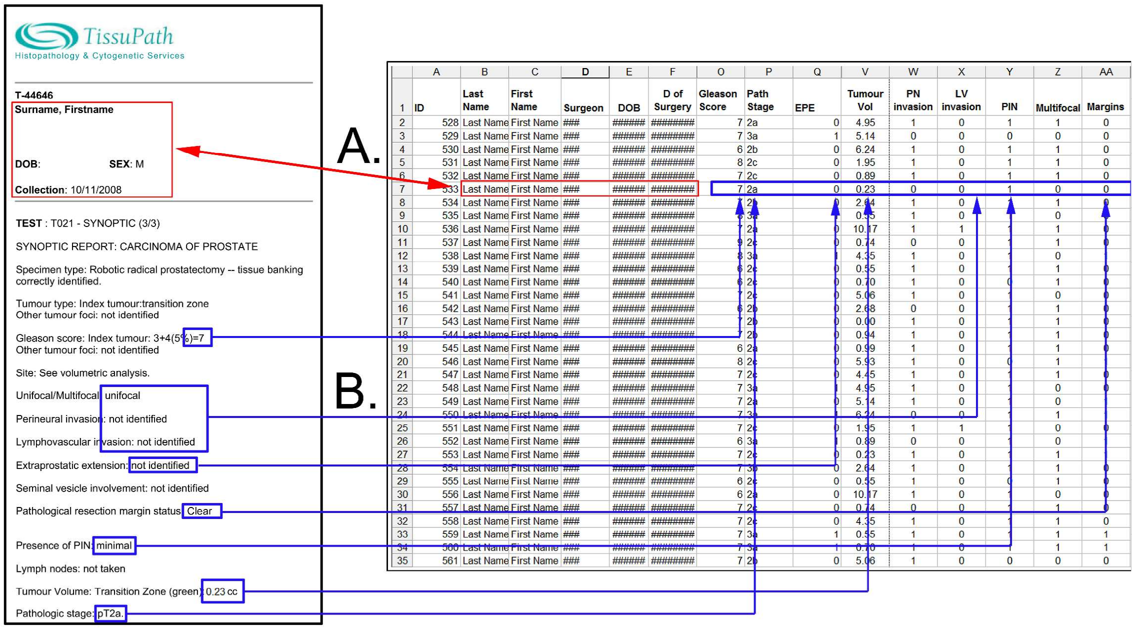
In this article we explored the reliability of manually transcribed pathology data across different fields in a large contemporary prostate cancer database. Initially housed as a Microsoft Excel spreadsheet, the database has evolved to become server based with a web-based interface. We have established automated electronic datafeeds from our pathology service provider to reduce the manual human data entry component in the pathology data, and we have used these to prospectively and retrospectively populate data fields. We compared overlapping data from the electronic feeds and previously manually entered data, whereby we could gauge the accuracy of pathology details in a subset of our patients. In this way, we could determine the error involved in manual data entry in different fields across patients with relative ease. In addition, we explored error that might be attributable to the source data.

PATIENT AND METHODS

Database Systems and Data Linking

Between 2004 and 2011, 1471 patients underwent radical prostatectomy across our institutions. Variables including demographics, pre-operative PSA, pathological Gleason score and stage, and other pathological data relating to the prostatectomy specimens were manually entered in a non-relational database (Microsoft Excel spreadsheet) for the first 853 of these cases with 57 fields per patient (2004-2008). Although most data collection was primarily prospectively performed, pathology data was obtained retrospectively once printed specimen pathology reports became available, or missing data was found on later review. For every patient, printed pathology reports were consulted and data manually transcribed into the spreadsheet. Each of the reports were issued by a single pathology group and consisted of one to two pages of prose. Since 2006, the reports have been accompanied by a separate page with a 'synoptic' report. This synoptic report contained the pathology data in a structured format with fields of interest listed on a single page enabling greater ease of interpretation over the traditional reports in prose. Several different junior medical staff performed manual data entry at sporadic times throughout. In 2010 our institution moved all data to Caisis 5.0, a web-based relational database system developed at the Memorial Sloan Kettering Cancer Centre in New York, and ceased manual recording of pathology data from hardcopy reports.

We subsequently established a data link between our database and the pathology group whereby electronically encrypted reports were provided in HL7 standard v2.31 format, a health industry information technology standard. The reports were retrieved using client/server software through a TCP/IP link. Custom software was developed in Visual Basic (Microsoft Visual Studio 2010) that enabled us to parse text or values of interest from the synoptic reports and automatically populate associated fields in our database system. (**Fig 1**). All new pathology data henceforth was imported in this way after testing of the new software for accuracy was performed. Digital import of data was conservative, and the software was written in such a way that data was not transferred in cases of ambiguity. Given that the original synoptic reports dating back to 2006 were digitally stored by our pathology group, we were able to import pathology data from this time period with our software and use it in preference to any data arising from manual data entry for patients from 2006 to 2011.



T-44646
Surname, Firstname

DOB: **SEX**: M

Collection: 10/11/2008

TEST : T021 - SYNOPTIC (3/3)

SYNOPTIC REPORT: CARCINOMA OF PROSTATE

Specimen type: Robotic radical prostatectomy -- tissue banking correctly identified.

Tumour type: Index tumour:transition zone
Other tumour foci: not identified

Gleason score: Index tumour: 3+4(5%)=7
Other tumour foci: not identified

Site: See volumetric analysis.

Unifocal/Multifocal: unifocal

Perineural invasion: not identified

Lymphovascular invasion: not identified

Extraprostatic extension: not identified

Seminal vesicle involvement: not identified

Pathological resection margin status: Clear

Presence of PIN: minimal

Lymph nodes: not taken

Tumour Volume: Transition Zone (green): 0.23 cc

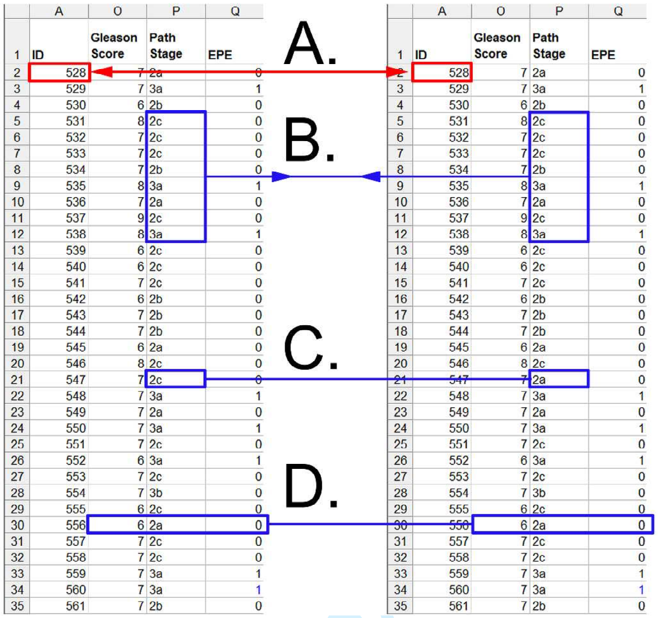
Pathologic stage: pT2a.

	A	B	C	D	E	F	O	P	Q	V	W	X	Y	Z	AA
1	ID	Last Name	First Name	Surgeon	DOB	D of Surgery	Gleason Score	Path Stage	EPE	Tumour Vol	PN invasion	LV invasion	PIN	Multifocal	Margins
2	528	Last Name	First Name	###	######	########	7	2a	0	4.95	1	0	1	1	0
3	529	Last Name	First Name	###	######	########	7	3a	1	5.14	0	0	0	0	0
4	530	Last Name	First Name	###	######	########	6	2b	0	6.24	1	0	1	1	0
5	531	Last Name	First Name	###	######	########	8	2c	0	1.95	1	0	1	1	0
6	532	Last Name	First Name	###	######	########	7	2c	0	0.89	1	0	1	1	0
7	533	Last Name	First Name	###	######	########	7	2a	0	0.23	0	0	1	0	0
8	534	Last Name	First Name	###	######	########	7	2b	0	2.64	1	0	1	1	0
9	535	Last Name	First Name	###	######	########	8	3a	1	0.55	1	0	1	0	[illegible]
10	536	Last Name	First Name	###	######	########	7	2a	0	10.17	1	1	1	1	0
11	537	Last Name	First Name	###	######	########	9	2c	0	0.74	0	0	1	1	0
12	538	Last Name	First Name	###	######	########	8	3a	1	4.35	1	0	1	0	1
13	539	Last Name	First Name	###	######	########	6	2c	0	0.55	1	0	1	1	0
14	540	Last Name	First Name	###	######	########	6	2c	0	0.70	1	0	0	1	0
15	541	Last Name	First Name	###	######	########	7	2c	0	5.06	1	0	1	0	0
16	542	Last Name	First Name	###	######	########	6	2b	0	2.68	0	0	1	1	0
17	543	Last Name	First Name	###	######	########	7	2b	0	0.00	1	0	1	1	0
18	544	Last Name	First Name	###	######	########	7	2b	0	0.94	1	0	1	1	0
19	545	Last Name	First Name	###	######	########	6	2a	0	0.99	1	0	1	1	0
20	546	Last Name	First Name	###	######	########	8	2c	0	5.93	1	0	0	1	1
21	547	Last Name	First Name	###	######	########	7	2c	0	4.45	1	0	1	1	0
22	548	Last Name	First Name	###	######	########	7	3a	1	4.95	1	0	1	0	0
23	549	Last Name	First Name	###	######	########	7	2a	0	5.14	1	0	1	0	1
24	550	Last Name	First Name	###	######	########	7	3a	1	6.24	0	0	1	1	1
25	551	Last Name	First Name	###	######	########	7	2c	0	1.95	1	1	1	0	0
26	552	Last Name	First Name	###	######	########	6	3a	1	0.89	0	0	1	0	1
27	553	Last Name	First Name	###	######	########	7	2c	0	0.23	1	0	1	1	1
28	554	Last Name	First Name	###	######	########	7	3b	0	2.64	1	0	1	1	0
29	555	Last Name	First Name	###	######	########	6	2c	0	0.55	1	0	0	1	0
30	556	Last Name	First Name	###	######	########	6	2a	0	10.17	1	0	1	0	0
31	557	Last Name	First Name	###	######	########	7	2c	0	0.74	0	0	1	1	0
32	558	Last Name	First Name	###	######	########	7	2c	0	4.35	1	0	1	1	0
33	559	Last Name	First Name	###	######	########	7	3a	1	0.55	1	0	1	1	1
34	560	Last Name	First Name	###	######	########	7	3a	1	0.70	1	0	1	1	1
35	561	Last Name	First Name	###	######	########	7	2b	0	5.06	1	0	0	0	0

Fig. 1 Schematic representation of the digital import of pathology data. Structured ‘synoptic’ reports facilitated digital recognition of relevant pathology fields. (A) Demographics data was used to link reports to individual patients in the database and (B) individual data were then extracted from the report and directed to populate relevant fields in the main database.

Analysis of Error in Manual Data Entry

The digitally imported data were placed parallel to any existing manual entry records and following institutional review board approval, we were able to directly compare between digitally imported data and manual entry data for 752 patients where records from the two data entry methods co-existed. We assumed that any mismatch in the fields within the manual entry and digital import was due to human error in the data entry, as data were copied from printed versions of these same reports in the first instance. The importing software had been extensively tested and errors would be systematic within each field rather than transcriptional in nature. We excluded from analysis specimen pathology fields with fewer than 200 comparable entries in order to detect at least a 0.5% error rate. 421 patients had at least 10 completed pathology fields in both manual entry and electronic import records and were thus selected for study. This would allow us to detect the error rate across a patient's fields and also minimise individual patient factors in explaining error rates in different fields. Within each pathology field, we linked records based on unique identifiers and electronically compared them using custom prepared software (Visual Basic, Microsoft Visual Studio 2010). We identified and counted any mismatches and then compared across all fields for each patient to determine the number of patients affected by one or more errors across the cohort (**Fig. 2**).



	A	O	P	Q
1	ID	Gleason Score	Path Stage	EPE
2	528	7	2a	0
3	529	7	3a	1
4	530	6	2b	0
5	531	8	2c	0
6	532	7	2c	0
7	533	7	2c	0
8	534	7	2b	0
9	535	8	3a	1
10	536	7	2a	0
11	537	9	2c	0
12	538	8	3a	1
13	539	6	2c	0
14	540	6	2c	0
15	541	7	2c	0
16	542	6	2b	0
17	543	7	2b	0
18	544	7	2b	0
19	545	6	2a	0
20	546	8	2c	0
21	547	7	2c	0
22	548	7	3a	1
23	549	7	2a	0
24	550	7	3a	1
25	551	7	2c	0
26	552	6	3a	1
27	553	7	2c	0
28	554	7	3b	0
29	555	6	2c	0
30	556	6	2a	0
31	557	7	2c	0
32	558	7	2c	0
33	559	7	3a	1
34	560	7	3a	1
35	561	7	2b	0

	A	O	P	Q
1	ID	Gleason Score	Path Stage	EPE
2	528	7	2a	0
3	529	7	3a	1
4	530	6	2b	0
5	531	8	2c	0
6	532	7	2c	0
7	533	7	2c	0
8	534	7	2b	0
9	535	8	3a	1
10	536	7	2a	0
11	537	9	2c	0
12	538	8	3a	1
13	539	6	2c	0
14	540	6	2c	0
15	541	7	2c	0
16	542	6	2b	0
17	543	7	2b	0
18	544	7	2b	0
19	545	6	2a	0
20	546	8	2c	0
21	547	7	2a	0
22	548	7	3a	1
23	549	7	2a	0
24	550	7	3a	1
25	551	7	2c	0
26	552	6	3a	1
27	553	7	2c	0
28	554	7	3b	0
29	555	6	2c	0
30	556	6	2a	0
31	557	7	2c	0
32	558	7	2c	0
33	559	7	3a	1
34	560	7	3a	1
35	561	7	2b	0

Fig. 2 Schematic representation of the comparison of a dataset imported digitally and in parallel to a manually entered dataset. (A) Records were linked using unique patient identifiers, and (B) pathology fields were individually compared. Concordant data were flagged for merging in order to eliminate duplicate data (C) Mismatches were used to identify errors in the manual entry dataset. (D) We compared across all pathology fields for individual patients.

Analysis of Error within Pathology Reports

In order to gain insight into the error inherent within the original pathology reports from which we sourced data, we measured the concordance between pathological stage and one of the descriptors that lead to stage determination, namely the extraprostatic extension variable. An error was detected in cases of incongruity where extraprostatic extension was present but the staging was T2, or extraprostatic extension was absent but the staging was T3. We identified and excluded the small number of cases of stage T3b disease where seminal vesicle invasion was apparent but extraprostatic extension not definite, as this would not be considered an error. We examined only the electronically imported data read directly from the synoptic reports, so that the effect of manual data

entry was excluded. We identified a total of 971 cases where both pathological stage and extraprostatic extension fields had both been successfully imported from the synoptic pathology reports (Table I). Once again using Visual Basic, we generated a report indicating cases where there was mismatch between pathological stage and extraprostatic extension status. The original cases in which these mismatches occurred were all reviewed by a pathologist to confirm the presence of error in the source material.

Table I

Pathological Stage	**Number**
T2a	131
T2b	4
T2c	543
T3a	225
T3b	68
T4	0
Extraprostatic Extension	
Absent	670
Present	302
Total Cases for Comparison	971

Statistical Analysis

Percentage error rates were calculated by dividing the absolute number of errors by the total number of data points examined overall and in each field. Binomial distribution was used to calculate 95% confidence intervals for these rates, and Fisher's Exact Test applied to 2x2 contingency tables where necessary (PASW Statistics 18.0; IBM, Chicago, Illinois, 2010). All statistical tests were two-tailed and significance was assumed at $\alpha<0.05$.

RESULTS

Of the 421 patients selected for this part of the study, 320 had completely concordant data between manual entry and electronic import methods in a median of 12 pathology fields (range 10-13), indicating an outright accuracy in 76% of all patients. Seventy one patients (16.8%) had errors in one field only, whilst 18 (4.3%) had two or more incorrect fields, and 12 (2.9%) had 3-5 errors.

Analysis of error rate in each individual pathology field yielded rates of error ranging from 0.5-6.4%. Across all fields, the error rate was 2.8% (Table II). Assuming that errors in different fields occurred independently of one another, that the fraction of records where at least one error occurred would be given by $1-(1-p)^n$, where p is the overall error rate and n is the number of fields. In this case, $1-(1-0.028)^{12} = 31\%$. Since the proportion of records where error occurred was 24%, the errors appear not to occur independently across the fields.

Fields involving descriptive parameters appeared more error-prone than those with direct measurements or involving numerical figures, so we grouped the fields based on data format. Of the 2658 data points involving numbers (numeric and alphanumeric), 30 (1.1%, 95% CI 0.78-1.6) were erroneous, compared with 116 (4.7%, 95% CI 3.9-5.6) of the 2490 data points with text (p<0.0001). The five fields that required an element of interpretation in data entry also appeared more error-prone and again, when data was pooled, their difference in error rates compared with fields allowing for direct transcription was significantly greater (5.2% vs 1.3%, p<0.0001).

Table II

Pathology Field	Variable Type	Data Format	Total Data Points	Error	Error Rate (%)	(95% CI)
Gleason 1	Categorical	Numeric	415	2	0.5%	(0.06-1.7)
Gleason 2	Categorical	Numeric	415	3	0.7%	(0.15-2.1)
Gleason score	Categorical	Numeric	415	1	0.2%	(0.01-1.3)
Extraprostatic Extension*	Binary	Text	421	21	5.0%	(3.1-7.5)
Stage	Categorical	Alphanumeric	421	13	3.1%	(1.7-5.2)
Focality	Binary	Text	421	9	2.1%	(1.0-4.0)
Perineural Invasion*	Categorical	Text	421	27	6.4%	(4.3-9.2)
Lymphovascular Invasion*	Categorical	Text	421	27	6.4%	(4.3-9.2)
Prostatic Intraepithelial Neoplasia*	Categorical	Text	420	27	6.4%	(4.3-9.2)
Margins*	Binary	Text	386	5	1.3%	(0.42-3.0)
Tumour Volume	Continuous	Numeric	310	4	1.3%	(0.35-3.3)
Prostate Dimensions+	Continuous	Numeric	272	2	0.7%	(0.09-2.6)
Prostate Weight	Continuous	Numeric	410	5	1.2%	(0.40-2.8)
All Fields			5148	146	2.8%	(2.4-3.3)

** Data required some interpretation on data entry – these were coded numerically*

+Each data point was a combination of 3 numbers. Error never occurred in more than one dimension.

In the 971 cases used for the analysis of source data error, six cases were staged T2 but in fact were positive for extraprostatic extension (6 of 672, 0.9%). On pathologist review, these cases had indeed been understaged. One case of a T3 prostate cancer erroneously stated on the synoptic report that extraprostatic extension was not identified (Table III). This was the sole inconsistency between the original prose pathology report and its accompanying synoptic report (1 of 971, 0.1%). Although only two variables have been analysed, these figures suggest a very low rate of baseline error inherent in the pathology reports.

Table III

Pathological Stage	Matches	Mismatches	Error Rate %	(95% CI)
T2	672	6	0.9%	(0.33-1.9)
T3	292	1	0.3%	(0.01-1.9)
Total	964	7	0.7%	(0.30-1.5)

DISCUSSION

In a large contemporary radical prostatectomy dataset we have examined pathology data in a subset of over 400 patients and found the overall error rate due to manual data entry to be 2.8% across all fields. Individual fields were found to vary in error rates between 0.5% and 6.4%, and those involving descriptive text or requiring an element of interpretation appeared more vulnerable to error. Almost a quarter of patients had at least one data error when all pathology fields were considered, as might occur when multivariable statistical analysis is undertaken. We have also examined the source data electronically without human influence and established a baseline error rate of less than 1%.

The strengths of our study include a combination of factors that enable a realistic representation of a small to moderate sized oncology database used for research purposes. As the data was stored in a simple spreadsheet, not collected for clinical use and was sourced from primary clinical documents, this context of data entry represents a common scenario predating modern informatics solutions. We also examined a distinct set of data fields with varying formats important to clinical oncological research. Together, these increase the relevance of our findings to cancer datasets in general, and in particular to data which has provenance in times prior to the introduction of more sophisticated modes of data entry. In addition, we have checked the integrity of one aspect of our source data, which is of importance in both clinical and academic settings. This helps to set a lower limit to the general error rate that can be achieved with interventions for data integrity imposed beyond initial pathology reporting.

Our study was limited by its use of a single spreadsheet from a single series of patients. Although different institutions may use different data systems, the maintenance of clinical datasets on such spreadsheets is common in the clinical environment. Our source data was in the form of synoptic reports designed for ease of data transcription, rather than traditional pathology reports in prose and this may have reduced the true error rate of such data. Other major limitations were in the study design, whereby we could not differentiate easily between different types of data entry error despite being able to infer this to some extent from the format of data. Due to the nature of spreadsheets, we could not definitely account for row or column shifts in blocks of data as a source of error, although, on visual inspection of the errors this did not seem to be the case. As we only examined pathology fields covered by electronic import, the findings were not representative of the entire dataset, which also includes operative and perioperative details, and thus the study was not designed to test the effect that these other factors may have had on error nor was it designed to detect errors in these important areas. A final limitation was that the size of error in fields containing continuous data was not measured as we only identified mismatches in the datasets, and this is required to assess more fully any impact of error in those fields.

Studies of error in clinical datasets are scarce, owing in part to the time and resources required to conduct these audits. Our overall error rate in manually entered data appears similar to that of previous studies. In one study occurring over 10 years, Zellner et al reported an estimated probability of error in two systems at about 2.4% and the estimated error frequency in a database alone was 2.7%[15]. In this case, less than 10% of the overall dataset was examined using random sampling. Arndt et al performed a

detailed study of observer rating scores in a multicentre field setting[11], whilst Goldberg et al examined several clinical research databases, with errors ranging from 2.3% to 26.9% detected by the double-entry method in fields relating to timepoints of disease and tumour recurrence status[7]. In general, these studies have involved more sophisticated data entry interfaces that allowed more detailed analysis of the underlying aetiology of data errors.

In contrast, our study has directly examined most fields in the subset of pathology variables, of particular importance in oncology research, and removed the effects of manual transfer of data in the generation of the comparison dataset. We also analysed error in the source data, as they might exceed those of data entry and render attempts to decrease downstream error frequency less meaningful. In this case the rate of 0.9% in mismatch between stage and extraprostatic extension was reassuringly lower than the overall manual entry dataset error frequency, and was also lower than the generally cited rate of 1.4% error for prostate pathology[16].

Although an analysis of the impact of data errors on outcomes was an area our study was unable to address, as follow up times were too short in our dataset for meaningful results, others have demonstrated the variable effects that erroneous data might have on outcomes such as rates of tumour recurrence and mortality rates[7 8]. While it is likely that a low rate of data error will have little effect in univariable analysis, studies involving many fields and demonstrating a small effect size with borderline significance levels are intuitively liable to the effect of errors. In these cases, and particularly where accuracy across a large feature space is essential such as in translational genomics research, it is

preferable that data errors are accounted for. Some investigators have developed corrective statistical tools to be used with a specified error rate in source databases in a particular circumstance[8], but such tools are unlikely to become widely applicable without more reporting of error rates within different types of datasets and analyses of outcome differences.

The greatest influence on error rates in our own clinical dataset was the transition to electronic data feeds from clinical sources and the application of software to retrospectively replace manually entered data. In doing so, we decreased the portion of patients with manually entered data from 58% (853/1471) to 9% (128/1471), although for various technical reasons many fields still remain manually inputted amongst the earlier patients. With advances in technology, it may even be possible to extract data from even earlier pathology reports, since all reports are typewritten, and maintain a dataset with virtually no manually entered pathology data. Where such manually entered data is unavoidable or forms part of a larger dataset, due acknowledgement of the provenance of data from different parts of that dataset by performing separate analysis or by employing sensitivity analysis might be considered in research. In addition, the judicious selection of pathology fields based on liability to error might be used.

The recognition that direct use of clinically attained data leads to better accuracy is not new. The need to re-key data between clinical sources and database interfaces has long been acknowledged to be a significant source of human error[17], and the removal of this aspect of data entry would presumably increase accuracy of clinical datasets overall. In recent times, the availability of data directly collected in the clinical setting for other

healthcare activities including medical research has increased. One clinical group in a peripheral hospital centralised data collection for audit and research purposes via web-browser based application software and extensively integrated the system in daily use. In just 12 months, their unit amassed over 3000 near complete patient records and reported enhanced accuracy due to the demonstration of the immediate clinical value of high-quality data capture to the users[18]. Such integrated record systems have been shown to have additional clinical benefits[19], whilst data collected as near in time and space as possible to the point of care is known to improve overall accuracy[20]. Indeed, pathologists currently generate pathology reports prospectively as part of the clinical process, and our use of electronic data feeds from our service provider is an example of how clinical data can be directly captured for clinical or research use without the need for manual data entry.

With increasing drive for the widespread implementation of electronic health records[21], comes the opportunity for more electronic data feeds into data repositories from health services such as those used in the present study, and this effectively diminishes the reliance on manual data entry. However, as our study demonstrated, even a clinical data source itself has a pervasive error rate, and there will remain a need for active error trapping. Furthermore, retrospective replacement of manually entered data may afford opportunities to examine the errors in manually entered data. Perhaps this work would allow the development of tools to adequately account for errors in early datasets that no technology can correct.

CONCLUSION

We have evaluated a large radical prostatectomy dataset and while the overall rate of error was low, individual pathology fields were variably prone to error. We have demonstrated the feasibility of checking source clinical data for error, and the possibility of attaining high quality data using electronic data feeds for both prospective and retrospective parts of our data repository.

ACKNOWLEDGEMENTS

Dr M.K.Hong was supported by scholarships from the Royal Australasian College of Surgeons and from the National Health and Medical Research Council, Australia. We gratefully acknowledge Roberto Leiva for his technical assistance.

CONFLICT OF INTEREST STATEMENT

No conflict of interest to declare.

CONTRIBUTORSHIP

Matthew KH Hong

Conceived and designed the study, interpreted the data, wrote manuscript

Henry HI Yao

Study design, data interpretation, critically reviewed article, revised article, final approval

John S Pedersen

Study design, Acquired and analysed data, critically reviewed article, final approval

Justin S Peters

Acquired data, critically reviewed article, final approval

Anthony J Costello

Acquired data, critically reviewed article, final approval

Declan G Murphy

Conception and design of study, critically reviewed article, final approval

Christopher M Hovens

Study design, data interpretation, revised article, final approval

Niall M Corcoran

Study design and conception, acquisition of data, interpretation of data, revised article, final approval

DATA SHARING

There is no additional data available **.**

FUNDING

None

References

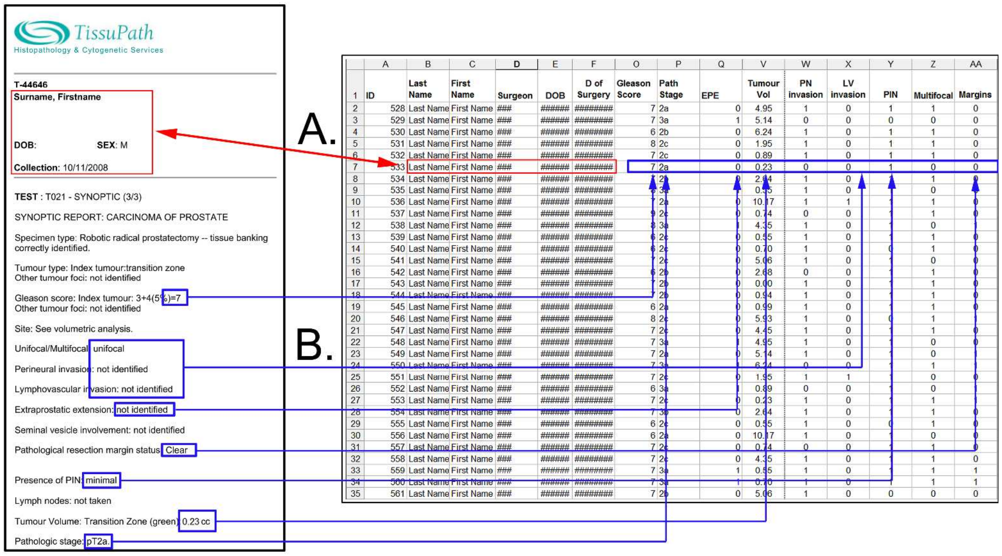
1. Hudson TJ, Anderson W, Artez A, Barker AD, Bell C, Bernabe RR, et al. International network of cancer genome projects. *Nature* 2010;464(7291):993-8.
2. Harel A, Dalah I, Pietrokovski S, Safran M, Lancet D. Omics data management and annotation. *Methods Mol Biol* 2011;719:71-96.
3. Needham DM, Sinopoli DJ, Dinglas VD, Berenholtz SM, Korupolu R, Watson SR, et al. Improving data quality control in quality improvement projects. *Int J Qual Health Care* 2009;21(2):145-50.
4. Arts DG, De Keizer NF, Scheffer GJ. Defining and improving data quality in medical registries: a literature review, case study, and generic framework. *J Am Med Inform Assoc* 2002;9(6):600-11.
5. Beretta L, Aldrovandi V, Grandi E, Citerio G, Stocchetti N. Improving the quality of data entry in a low-budget head injury database. *Acta Neurochir (Wien)* 2007;149(9):903-9.
6. Seddon DJ, Williams EM. Data quality in population-based cancer registration: an assessment of the Merseyside and Cheshire Cancer Registry. *Br J Cancer* 1997;76(5):667-74.
7. Goldberg SI, Niemierko A, Turchin A. Analysis of data errors in clinical research databases. *AMIA Annu Symp Proc* 2008:242-6.
8. Gallivan S, Pagel C. Modelling of errors in databases. *Health Care Management Science* 2008;11(1):35-40.
9. Goldberg SI, Niemierko A, Shubina M, Turchin A. "Summary Page": a novel tool that reduces omitted data in research databases. *BMC Med Res Methodol* 2010;10:91.
10. Warsi AA, White S, McCulloch P. Completeness of data entry in three cancer surgery databases. *Eur J Surg Oncol* 2002;28(8):850-6.
11. Arndt S, Tyrrell G, Woolson RF, Flaum M, Andreasen NC. Effects of errors in a multicenter medical study: preventing misinterpreted data. *J Psychiatr Res* 1994;28(5):447-59.
12. Hayrinen K, Saranto K, Nykanen P. Definition, structure, content, use and impacts of electronic health records: a review of the research literature. *Int J Med Inform* 2008;77(5):291-304.
13. Antonarakis ES, Feng Z, Trock BJ, Humphreys EB, Carducci MA, Partin AW, et al. The natural history of metastatic progression in men with prostate-specific antigen recurrence after radical prostatectomy: long-term follow-up. *BJU Int* 2012;109(1):32-9.
14. Neo EL, Beeke C, Price T, Maddern G, Karapetis C, Luke C, et al. South Australian clinical registry for metastatic colorectal cancer. *ANZ J Surg* 2011;81(5):352-7.
15. Zellner D, Schromm T, Frankewitsch T, Giehl M, Keller F. Structured data entry for reliable acquisition of pharmacokinetic data. *Methods Inf Med* 1996;35(3):261-4.
16. Frable WJ. Surgical pathology--second reviews, institutional reviews, audits, and correlations: what's out there? Error or diagnostic variation? *Arch Pathol Lab Med* 2006;130(5):620-5.
17. Gostel R. HyperCard to SPSS: improving data integrity. *Comput Nurs* 1993;11(1):25-8.

18. Tran P, Morrison SG, Lade JA, Haw CS. An integrated approach to surgical audit. *ANZ J Surg* 2011;81(5):313-4.

19. Featherstone I, Keen J. Do integrated record systems lead to integrated services? An observational study of a multi-professional system in a diabetes service. *Int J Med Inform* 2012;81(1):45-52.

20. Veen EJ, Janssen-Heijnen ML, Leenen LP, Roukema JA. The registration of complications in surgery: a learning curve. *World J Surg* 2005;29(3):402-9.

21. Pearce C, Haikerwal MC. E-health in Australia: time to plunge into the 21st century. *Med J Aust* 2010;193(7):397-8.



TissuPath
Histopathology & Cytogenetic Services

T-44646
Surname, Firstname

DOB: SEX: M

Collection: 10/11/2008

TEST : T021 - SYNOPTIC (3/3)

SYNOPTIC REPORT: CARCINOMA OF PROSTATE

Specimen type: Robotic radical prostatectomy -- tissue banking correctly identified.

Tumour type: Index tumour:transition zone
Other tumour foci: not identified

Gleason score: Index tumour: 3+4(5%)=7
Other tumour foci: not identified

Site: See volumetric analysis.

Unifocal/Multifocal: unifocal

Perineural invasion: not identified

Lymphovascular invasion: not identified

Extraprostatic extension: not identified

Seminal vesicle involvement: not identified

Pathological resection margin status: Clear

Presence of PIN: minimal

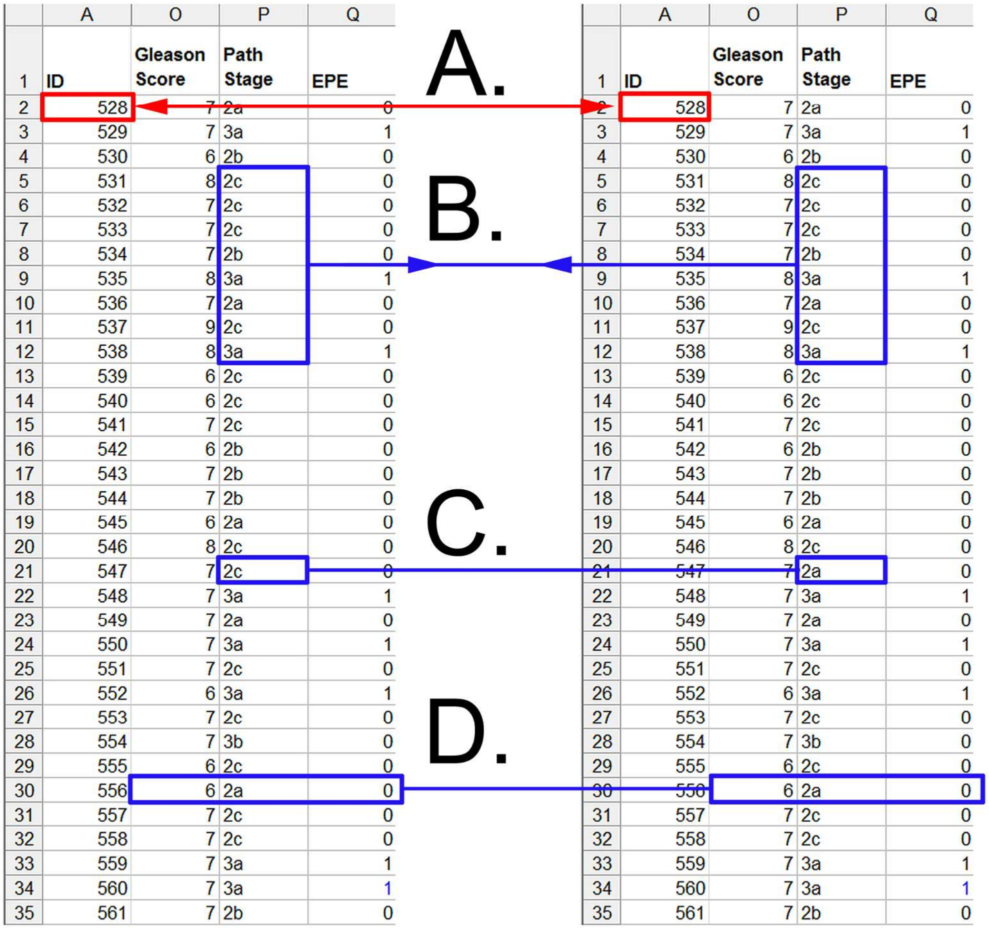
Lymph nodes: not taken

Tumour Volume: Transition Zone (green): 0.23 cc

Pathologic stage: pT2a.

	A	B	C	D	E	F	O	P	Q	V	W	X	Y	Z	AA
1	ID	Last Name	First Name	Surgeon	DOB	D of Surgery	Gleason Score	Path Stage	EPE	Tumour Vol	PN invasion	LV invasion	PIN	Multifocal	Margins
2	528	Last Name	First Name	###	######	########	7	2a	0	4.95	1	0	1	1	0
3	529	Last Name	First Name	###	######	########	7	3a	1	5.14	0	0	0	0	0
4	530	Last Name	First Name	###	######	########	6	2b	0	6.24	1	0	1	1	0
5	531	Last Name	First Name	###	######	########	8	2c	0	1.95	1	0	1	1	0
6	532	Last Name	First Name	###	######	########	7	2c	0	0.89	1	0	1	1	0
7	533	Last Name	First Name	###	######	########	7	2a	0	0.23	0	0	1	0	0
8	534	Last Name	First Name	###	######	########	7	2b	0	2.64	1	0	[illegible]	1	0
9	535	Last Name	First Name	###	######	########	8	3a	[illegible]	0.55	1	0	[illegible]	0	[illegible]
10	536	Last Name	First Name	###	######	########	7	2a	0	10.17	1	1	1	1	0
11	537	Last Name	First Name	###	######	########	9	2c	0	0.74	0	0	1	1	0
12	538	Last Name	First Name	###	######	########	8	3a	1	4.35	1	0	1	0	1
13	539	Last Name	First Name	###	######	########	6	2c	0	0.55	1	0	1	1	0
14	540	Last Name	First Name	###	######	########	6	2c	0	0.70	1	0	0	1	0
15	541	Last Name	First Name	###	######	########	7	2c	0	5.06	1	0	1	0	0
16	542	Last Name	First Name	###	######	########	6	2b	0	2.68	0	0	1	1	0
17	543	Last Name	First Name	###	######	########	7	2b	0	0.00	1	0	1	1	0
18	544	Last Name	First Name	###	######	########	7	2b	0	0.94	1	0	1	1	0
19	545	Last Name	First Name	###	######	########	6	2a	0	0.99	1	0	1	1	0
20	546	Last Name	First Name	###	######	########	8	2c	0	5.93	1	0	0	1	1
21	547	Last Name	First Name	###	######	########	7	2c	0	4.45	1	0	1	1	0
22	548	Last Name	First Name	###	######	########	7	3a	1	4.95	1	0	1	0	0
23	549	Last Name	First Name	###	######	########	7	2a	0	5.14	1	0	1	0	1
24	550	Last Name	First Name	###	######	########	7	3a	1	6.24	0	0	1	1	1
25	551	Last Name	First Name	###	######	########	7	2c	0	1.95	1	1	1	0	0
26	552	Last Name	First Name	###	######	########	6	3a	1	0.89	0	0	1	0	1
27	553	Last Name	First Name	###	######	########	7	2c	0	0.23	1	0	1	1	1
28	554	Last Name	First Name	###	######	########	7	3b	0	2.64	1	0	1	1	0
29	555	Last Name	First Name	###	######	########	6	2c	0	0.55	1	0	0	1	0
30	556	Last Name	First Name	###	######	########	6	2a	0	10.17	1	0	1	0	0
31	557	Last Name	First Name	###	######	########	7	2c	0	0.74	0	0	1	1	0
32	558	Last Name	First Name	###	######	########	7	2c	0	4.35	1	0	1	1	0
33	559	Last Name	First Name	###	######	########	7	3a	1	0.55	1	0	1	1	1
34	560	Last Name	First Name	###	######	########	7	3a	1	0.70	1	0	1	1	1
35	561	Last Name	First Name	###	######	########	7	2b	0	5.06	1	0	0	0	0

Fig. 1 Schematic representation of the digital import of pathology data. Structured 'synoptic' reports facilitated digital recognition of relevant pathology fields. (A) Demographics data was used to link reports to individual patients in the database and (B) individual data were then extracted from the report and directed to populate relevant fields in the main database.
389x223mm (300 x 300 DPI)



	A	O	P	Q
1	ID	Gleason Score	Path Stage	EPE
2	528	7	2a	0
3	529	7	3a	1
4	530	6	2b	0
5	531	8	2c	0
6	532	7	2c	0
7	533	7	2c	0
8	534	7	2b	0
9	535	8	3a	1
10	536	7	2a	0
11	537	9	2c	0
12	538	8	3a	1
13	539	6	2c	0
14	540	6	2c	0
15	541	7	2c	0
16	542	6	2b	0
17	543	7	2b	0
18	544	7	2b	0
19	545	6	2a	0
20	546	8	2c	0
21	547	7	2c	0
22	548	7	3a	1
23	549	7	2a	0
24	550	7	3a	1
25	551	7	2c	0
26	552	6	3a	1
27	553	7	2c	0
28	554	7	3b	0
29	555	6	2c	0
30	556	6	2a	0
31	557	7	2c	0
32	558	7	2c	0
33	559	7	3a	1
34	560	7	3a	1
35	561	7	2b	0

	A	O	P	Q
1	ID	Gleason Score	Path Stage	EPE
2	528	7	2a	0
3	529	7	3a	1
4	530	6	2b	0
5	531	8	2c	0
6	532	7	2c	0
7	533	7	2c	0
8	534	7	2b	0
9	535	8	3a	1
10	536	7	2a	0
11	537	9	2c	0
12	538	8	3a	1
13	539	6	2c	0
14	540	6	2c	0
15	541	7	2c	0
16	542	6	2b	0
17	543	7	2b	0
18	544	7	2b	0
19	545	6	2a	0
20	546	8	2c	0
21	547	7	2a	0
22	548	7	3a	1
23	549	7	2a	0
24	550	7	3a	1
25	551	7	2c	0
26	552	6	3a	1
27	553	7	2c	0
28	554	7	3b	0
29	555	6	2c	0
30	556	6	2a	0
31	557	7	2c	0
32	558	7	2c	0
33	559	7	3a	1
34	560	7	3a	1
35	561	7	2b	0

Fig. 2 Schematic representation of the comparison of a dataset imported digitally and in parallel to a manually entered dataset. (A) Records were linked using unique patient identifiers, and (B) pathology fields were individually compared. Concordant data were flagged for merging in order to eliminate duplicate data (C) Mismatches were used to identify errors in the manual entry dataset. (D) We compared across all pathology fields for individual patients.
86x79mm (600 x 600 DPI)

Table I

Pathological Stage	Number
T2a	131
T2b	4
T2c	543
T3a	225
T3b	68
T4	0
Extraprostatic Extension	
Absent	670
Present	302
Total Cases for Comparison	971

Table II

Pathology Field	Variable Type	Data Format	Total Data Points	Error	Error Rate (%)	(95% CI)
Gleason 1	Categorical	Numeric	415	2	0.5%	(0.06-1.7)
Gleason 2	Categorical	Numeric	415	3	0.7%	(0.15-2.1)
Gleason score	Categorical	Numeric	415	1	0.2%	(0.01-1.3)
Extraprostatic Extension*	Binary	Text	421	21	5.0%	(3.1-7.5)
Stage	Categorical	Alphanumeric	421	13	3.1%	(1.7-5.2)
Focality	Binary	Text	421	9	2.1%	(1.0-4.0)
Perineural Invasion*	Categorical	Text	421	27	6.4%	(4.3-9.2)
Lymphovascular Invasion*	Categorical	Text	421	27	6.4%	(4.3-9.2)
Prostatic Intraepithelial Neoplasia*	Categorical	Text	420	27	6.4%	(4.3-9.2)
Margins*	Binary	Text	386	5	1.3%	(0.42-3.0)
Tumour Volume	Continuous	Numeric	310	4	1.3%	(0.35-3.3)
Prostate Dimensions+	Continuous	Numeric	272	2	0.7%	(0.09-2.6)
Prostate Weight	Continuous	Numeric	410	5	1.2%	(0.40-2.8)
All Fields			5148	146	2.8%	(2.4-3.3)

** Data required some interpretation on data entry – these were coded numerically*

+Each data point was a combination of 3 numbers. Error never occurred in more than one dimension.

Table III

Pathological Stage	Matches	Mismatches	Error Rate %	(95% CI)
T2	672	6	0.9%	(0.33-1.9)
T3	292	1	0.3%	(0.01-1.9)
Total	964	7	0.7%	(0.30-1.5)

Error Rates in a Clinical Data Repository: Lessons from the Transition to Electronic Data Transfer

Journal:	*BMJ Open*
Manuscript ID:	bmjopen-2012-002406.R1
Article Type:	Research
Date Submitted by the Author:	09-Apr-2013
Complete List of Authors:	Hong, Matthew KH; Royal Melbourne Hospital, Department of Surgery Yao, Henry Han-I; Royal Melbourne Hospital, Department of Surgery Pedersen, John; TissuPath Specialist Pathology, Peters, Justin; Royal Melbourne Hospital, Department of Urology Costello, Anthony; Royal Melbourne Hospital, Department of Urology Murphy, Declan; Peter MacCallum Cancer Centre, Urological Service Team Hovens, Chris; Royal Melbourne Hospital, Department of Surgery Corcoran, Niall; Royal Melbourne Hospital, Department of Urology
<b>Primary Subject Heading</b>:	Health informatics
Secondary Subject Heading:	Urology, Oncology, Pathology
Keywords:	database, prostate cancer, data quality, error sources, clinical informatics

Error Rates in a Clinical Data Repository: Lessons from the Transition to Electronic Data Transfer

Matthew KH Hong
Henry HI Yao
John S Pedersen[*]
Justin S Peters
Anthony J Costello
Declan G Murphy[+]
Christopher M Hovens
Niall M Corcoran

Division of Urology, Department of Surgery, University of Melbourne, Royal Melbourne Hospital and the Australian Prostate Cancer Research Centre Epworth, Victoria, Australia.
* TissuPath Specialist Pathology, Mount Waverley and Monash University Faculty of Medicine, Victoria, Australia
+ Peter MacCallum Cancer Centre, East Melbourne, Australia

Correspondence:
Dr Matthew Hong,
Department of Surgery
University of Melbourne
Royal Melbourne Hospital
Grattan St Parkville
VIC 3050 Australia
T: 613 9342 7703
E: m.k.hong@ausdoctors.net

Keywords: Database, prostate cancer, data quality, error sources, clinical informatics
Word count: 3927 words

Article Summary

Article focus

- Although use of structured electronic databases is widespread, a substantial amount of clinical data used in research predates this.
- There is a paucity of literature on error rates in such clinical datasets used in research.
- We explored the reliability of manually transcribed data across different pathology fields in a prostate cancer database and also measured error rates attributable to the source data.

Key messages

- Whilst overall rate of error for manually entered data can be low, individual fields may be variably prone to error, especially those involving descriptive text or requiring an element of interpretation.
- Computerised systems can be used to check clinical source data for error.
- The use of electronic data feeds retrospectively can replace manually collected data fields in some cases to improve overall accuracy.

Strengths and limitations of this study

- Our study design provides a realistic representation of a small to moderate sized oncology database used for research purposes.
- We checked the integrity of one aspect of our source data.
- Our study was limited by its use of a single spreadsheet from a single series of patients.

- As we only examined pathology fields covered by electronic import, the findings were not representative of the entire dataset.

ABSTRACT

Objective: Data errors are a well-documented part of clinical datasets as is their potential to confound downstream analysis. In this study we explore the reliability of manually transcribed data across different pathology fields in a prostate cancer database and also measure error rates attributable to the source data.

Design: Descriptive study

Setting: Specialist urology service at a single centre in metropolitan Victoria in Australia

Participants: Between 2004 and 2011, 1471 patients underwent radical prostatectomy at our institution. In a large proportion of these cases, clinicopathological variables were recorded by manual data-entry. In 2011, we obtained electronic versions of the same printed pathology reports for our cohort. The data were electronically imported in parallel to any existing manual entry record enabling direct comparison between them.

Outcome measures: Error rates of manually entered data compared with electronically imported data across clinicopathological fields.

Results: 421 patients had at least 10 comparable pathology fields between the electronic import and manual records and were selected for study. 320 patients had concordant data between manually entered and electronically populated fields in a median of 12 pathology fields (range 10-13), indicating an outright accuracy in manually entered pathology data in 76% of patients. Across all fields, the error rate was 2.8% whilst individual field error ranges from 0.5-6.4%. Fields in text formats were significantly more error-prone than those with direct measurements or involving numerical figures (p<0.001). 971 cases were available for review of error within the source data, with figures of 0.1%-0.9%.

Conclusion: While the overall rate of error was low in manually entered data, individual pathology fields were variably prone to error. High quality pathology data can be obtained for both prospective and retrospective parts of our data repository and the electronic checking of source pathology data for error is feasible.

BACKGROUND AND SIGNIFICANCE

The majority of clinical research publications are based on the analysis of prospectively collected, clinical databases. In addition, patient centred databases are increasingly important in translational research efforts, as appropriately annotated tissue banks are the foundation for global multi-institutional collaborative efforts in genetic and epigenetic screening of various diseases[1]. Yet despite the stringent quality controls placed on the vast amounts of research data derived from these studies and the acute awareness of the need to control data quality[2 3], the inherent accuracy of original clinical datasets is one area that receives relatively little attention.

Data errors are common in clinical datasets[4-6], with some cancer databases recording error rates as high as almost 27% in some fields [7]. Such errors have the potential to adversely affect data analysis and interpretation, and can lead to erroneous conclusions[8]. Methods to first identify then correct errors in these datasets would be immensely valuable in the setting of the large-scale genomics projects being performed.

Two types of errors are described in the literature: one of omission, and one of erroneous value. Although it is sometimes argued that missing values carry greater impact due to their greater prevalence[9], which may be up to 55% in cancer surgery databases[10], these errors are more easily detected with judicious computer queries and corrected with retrospective data collection. On the contrary, once erroneous values permeate a dataset, their effects can cascade in unpredictable ways. Errors in high impact fields have been shown to adversely affect the interpretation of statistical analyses, even if the errors are at low prevalence[11]. Whilst it is well known that

structured data entry improves the accuracy of manual documentation[12], much of the clinical data of high value to researchers predates any effective informatics solutions aimed at data quality that might exist today. Instead, manual retrospective transcription of data from clinical records into relatively unstructured spreadsheets constitutes the data entry method for many clinical audits that subsequently serve research purposes. These datasets may have even transitioned to more carefully constructed data entry interfaces, as might occur in conditions such as prostate cancer where long follow up times of over ten years are necessary for study of oncological outcomes[13]. In such cases, the provenance of the data collected with earlier means may not be accounted for with subsequent analysis.

Studies involving large cancer datasets rarely report error rates or their management, and it is difficult to assess the impact that these may have on the outcomes reported[14]. Given the considerable effort that generally goes towards the collection of data for a large clinical database, it is unsurprising that surplus resources are usually unavailable to place towards the check of data accuracy. Although larger numbers in databases may be used to counter the problem of errors, the combination of datasets, particularly with different fields would serve only to magnify error rates.

Knowledge of errors in manually collected data could give insight into how these may be accounted for in subsequent analysis. In cancer databases, pathology data are of particular importance as they are relied upon to build cohorts of clinical relevance in research. Often there are multiple fields that give equivalent indications of underlying biology and any one could be analysed to similar effect. For example, both the

percentage involvement of tumour or its measured size may serve as parameters of cancer burden. It is unlikely that different types of data would have equivalent vulnerability to error, and knowledge of the fields or types of fields that might be more error-prone with manual data entry could help researchers judiciously select fields for analysis based on greater accuracy. It might also aid informaticians to focus on error prevention in fields that carry particular importance in clinical and research settings. In addition, it is important to measure the baseline level of error inherent within the pathology reports themselves, the data source, as no degree of accuracy in manual transcription or even automated processes can result in a lower error rate without amendment of the original report.

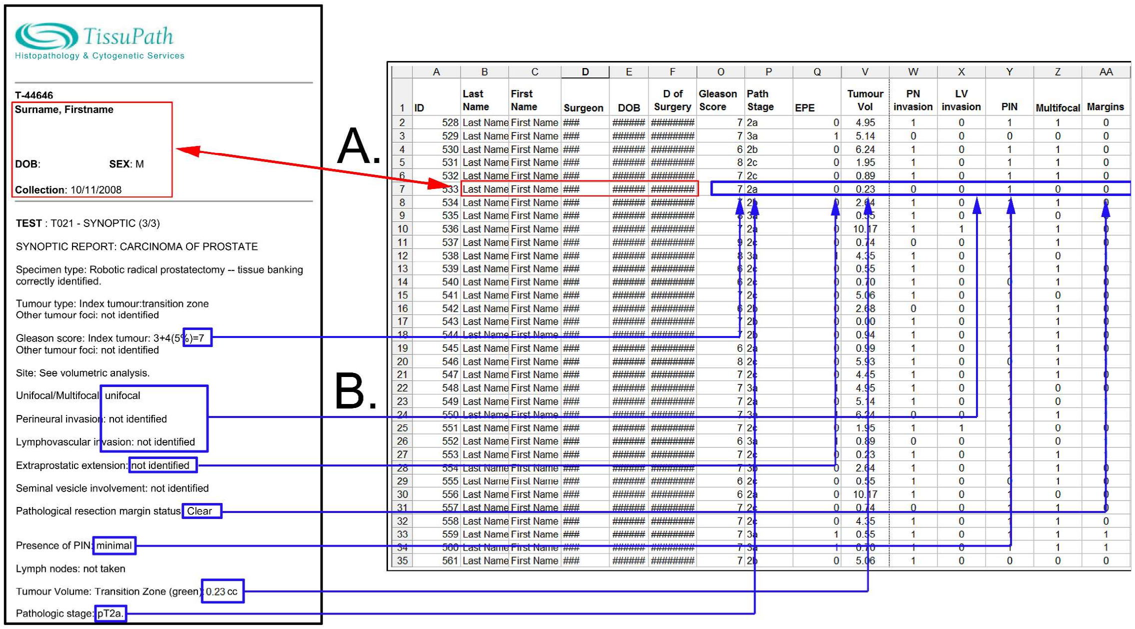
In this article we explored the reliability of manually transcribed pathology data across different fields in a large contemporary prostate cancer database. Initially housed as a Microsoft Excel spreadsheet, the database has evolved to become server based with a web-based interface. We have established automated electronic datafeeds from our pathology service provider to reduce the manual human data entry component in the pathology data, and we have used these to prospectively and retrospectively populate data fields. We compared overlapping data from the electronic feeds and previously manually entered data, whereby we could gauge the accuracy of pathology details in a subset of our patients. In this way, we could determine the error involved in manual data entry in different fields across patients with relative ease. In addition, we explored error that might be attributable to the source data.

PATIENT AND METHODS

Database Systems and Data Linking

Between 2004 and 2011, 1471 patients underwent radical prostatectomy across our institutions. Variables including demographics, pre-operative PSA, pathological Gleason score and stage, and other pathological data relating to the prostatectomy specimens were manually entered in a non-relational database (Microsoft Excel spreadsheet) for the first 853 of these cases with 57 fields per patient (2004-2008). Although most data collection was primarily prospectively performed, pathology data was obtained retrospectively once printed specimen pathology reports became available, or missing data was found on later review. For every patient, printed pathology reports were consulted and data manually transcribed into the spreadsheet. Each of the reports were issued by a single pathology group and consisted of one to two pages of prose. Since 2006, the reports have been accompanied by a separate page with a 'synoptic' report. This synoptic report contained the pathology data in a structured format with fields of interest listed on a single page enabling greater ease of interpretation over the traditional reports in prose. Manual data entry was performed by four surgical residents with knowledge of prostate cancer pathology and versed in the relevant terminology. In 2010 our institution moved all data to Caisis 5.0, a web-based relational database system developed at the Memorial Sloan Kettering Cancer Centre in New York, and ceased manual recording of pathology data from hardcopy reports.

We subsequently established a data link between our database and the pathology group whereby electronically encrypted reports were provided in HL7 standard v2.31 format, a health industry information technology standard. The reports were retrieved using client/server software through a TCP/IP link. Custom software was developed in Visual Basic (Microsoft Visual Studio 2010) that enabled us to parse text or values of interest from the synoptic reports and automatically populate associated fields in our database system. (**Fig 1**). All new pathology data henceforth was imported in this way after testing of the new software for accuracy was performed. Digital import of data was conservative, and the software was written in such a way that data was not transferred in cases of ambiguity. Given that the original synoptic reports dating back to 2006 were digitally stored by our pathology group, we were able to import pathology data from this time period with our software and use it in preference to any data arising from manual data entry for patients from 2006 to 2011.



TissuPath
Histopathology & Cytogenetic Services

T-44646
Surname, Firstname

DOB: **SEX**: M

Collection: 10/11/2008

TEST : T021 - SYNOPTIC (3/3)

SYNOPTIC REPORT: CARCINOMA OF PROSTATE

Specimen type: Robotic radical prostatectomy -- tissue banking correctly identified.

Tumour type: Index tumour:transition zone
Other tumour foci: not identified

Gleason score: Index tumour: 3+4(5%)=7
Other tumour foci: not identified

Site: See volumetric analysis.

Unifocal/Multifocal: unifocal

Perineural invasion: not identified

Lymphovascular invasion: not identified

Extraprostatic extension: not identified

Seminal vesicle involvement: not identified

Pathological resection margin status: Clear

Presence of PIN: minimal

Lymph nodes: not taken

Tumour Volume: Transition Zone (green): 0.23 cc

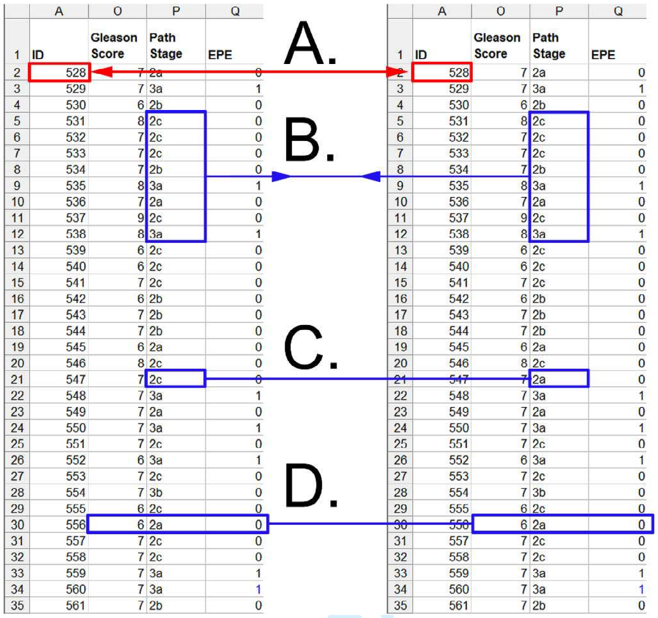
Pathologic stage: pT2a.

	A	B	C	D	E	F	O	P	Q	V	W	X	Y	Z	AA
1	ID	Last Name	First Name	Surgeon	DOB	D of Surgery	Gleason Score	Path Stage	EPE	Tumour Vol	PN invasion	LV invasion	PIN	Multifocal	Margins
2	528	Last Name	First Name	###	######	########	7	2a	0	4.95	1	0	1	1	0
3	529	Last Name	First Name	###	######	########	7	3a	1	5.14	0	0	0	0	0
4	530	Last Name	First Name	###	######	########	6	2b	0	6.24	1	0	1	1	0
5	531	Last Name	First Name	###	######	########	8	2c	0	1.95	1	0	1	1	0
6	532	Last Name	First Name	###	######	########	7	2c	0	0.89	1	0	1	1	0
7	533	Last Name	First Name	###	######	########	7	2a	0	0.23	0	0	1	0	0
8	534	Last Name	First Name	###	######	########	7	2b	0	2.64	1	0	1	1	0
9	535	Last Name	First Name	###	######	########	8	3a	1	0.55	1	0	1	0	0
10	536	Last Name	First Name	###	######	########	7	2a	0	10.17	1	1	1	1	0
11	537	Last Name	First Name	###	######	########	9	2c	0	0.74	0	0	1	1	0
12	538	Last Name	First Name	###	######	########	8	3a	1	4.35	1	0	1	0	1
13	539	Last Name	First Name	###	######	########	6	2c	0	0.55	1	0	1	1	0
14	540	Last Name	First Name	###	######	########	6	2c	0	0.70	1	0	0	1	0
15	541	Last Name	First Name	###	######	########	7	2c	0	5.06	1	0	1	0	0
16	542	Last Name	First Name	###	######	########	6	2b	0	2.68	0	0	1	1	0
17	543	Last Name	First Name	###	######	########	7	2b	0	0.00	1	0	1	1	0
18	544	Last Name	First Name	###	######	########	7	2b	0	0.94	1	0	1	1	0
19	545	Last Name	First Name	###	######	########	6	2a	0	0.99	1	0	1	1	0
20	546	Last Name	First Name	###	######	########	8	2c	0	5.93	1	0	0	1	1
21	547	Last Name	First Name	###	######	########	7	2c	0	4.45	1	0	1	1	0
22	548	Last Name	First Name	###	######	########	7	3a	1	4.95	1	0	1	0	0
23	549	Last Name	First Name	###	######	########	7	2a	0	5.14	1	0	1	0	1
24	550	Last Name	First Name	###	######	########	7	3a	1	6.24	0	0	1	1	1
25	551	Last Name	First Name	###	######	########	7	2c	0	1.95	1	1	1	0	0
26	552	Last Name	First Name	###	######	########	6	3a	1	0.89	0	0	1	0	1
27	553	Last Name	First Name	###	######	########	7	2c	0	0.23	1	0	1	1	1
28	554	Last Name	First Name	###	######	########	7	3b	0	2.64	1	0	1	1	0
29	555	Last Name	First Name	###	######	########	6	2c	0	0.55	1	0	0	1	0
30	556	Last Name	First Name	###	######	########	6	2a	0	10.17	1	0	1	0	0
31	557	Last Name	First Name	###	######	########	7	2c	0	0.74	0	0	1	1	0
32	558	Last Name	First Name	###	######	########	7	2c	0	4.35	1	0	1	1	0
33	559	Last Name	First Name	###	######	########	7	3a	1	0.55	1	0	1	1	1
34	560	Last Name	First Name	###	######	########	7	3a	1	0.70	1	0	1	1	1
35	561	Last Name	First Name	###	######	########	7	2b	0	5.06	1	0	0	0	0

Fig. 1 Schematic representation of the digital import of pathology data. Structured 'synoptic' reports facilitated digital recognition of relevant pathology fields. (A) Demographics data was used to link reports to individual patients in the database and (B) individual data were then extracted from the report and directed to populate relevant fields in the main database.

Analysis of Error in Manual Data Entry

The digitally imported data were placed parallel to any existing manual entry records and following institutional review board approval, we were able to directly compare between digitally imported data and manual entry data for 752 patients where records from the two data entry methods co-existed. We assumed that any mismatch in the fields within the manual entry and digital import was due to human error in the data entry, as data were copied from printed versions of these same reports in the first instance. The importing software had been extensively tested and errors would be systematic within each field rather than transcriptional in nature. We excluded from analysis specimen pathology fields with fewer than 200 comparable entries in order to detect at least a 0.5% error rate. 421 patients had at least 10 completed pathology fields in both manual entry and electronic import records and were thus selected for study. This would allow us to detect the error rate across a patient's fields and also minimise individual patient factors in explaining error rates in different fields. Within each pathology field, we linked records based on unique identifiers and electronically compared them using custom prepared software (Visual Basic, Microsoft Visual Studio 2010). We identified and counted any mismatches and then compared across all fields for each patient to determine the number of patients affected by one or more errors across the cohort (**Fig. 2**).



	A	O	P	Q
1	ID	Gleason Score	Path Stage	EPE
2	528	7	2a	0
3	529	7	3a	1
4	530	6	2b	0
5	531	8	2c	0
6	532	7	2c	0
7	533	7	2c	0
8	534	7	2b	0
9	535	8	3a	1
10	536	7	2a	0
11	537	9	2c	0
12	538	8	3a	1
13	539	6	2c	0
14	540	6	2c	0
15	541	7	2c	0
16	542	6	2b	0
17	543	7	2b	0
18	544	7	2b	0
19	545	6	2a	0
20	546	8	2c	0
21	547	7	2c	0
22	548	7	3a	1
23	549	7	2a	0
24	550	7	3a	1
25	551	7	2c	0
26	552	6	3a	1
27	553	7	2c	0
28	554	7	3b	0
29	555	6	2c	0
30	556	6	2a	0
31	557	7	2c	0
32	558	7	2c	0
33	559	7	3a	1
34	560	7	3a	1
35	561	7	2b	0

	A	O	P	Q
1	ID	Gleason Score	Path Stage	EPE
2	528	7	2a	0
3	529	7	3a	1
4	530	6	2b	0
5	531	8	2c	0
6	532	7	2c	0
7	533	7	2c	0
8	534	7	2b	0
9	535	8	3a	1
10	536	7	2a	0
11	537	9	2c	0
12	538	8	3a	1
13	539	6	2c	0
14	540	6	2c	0
15	541	7	2c	0
16	542	6	2b	0
17	543	7	2b	0
18	544	7	2b	0
19	545	6	2a	0
20	546	8	2c	0
21	547	7	2a	0
22	548	7	3a	1
23	549	7	2a	0
24	550	7	3a	1
25	551	7	2c	0
26	552	6	3a	1
27	553	7	2c	0
28	554	7	3b	0
29	555	6	2c	0
30	556	6	2a	0
31	557	7	2c	0
32	558	7	2c	0
33	559	7	3a	1
34	560	7	3a	1
35	561	7	2b	0

Fig. 2 Schematic representation of the comparison of a dataset imported digitally and in parallel to a manually entered dataset. (A) Records were linked using unique patient identifiers, and (B) pathology fields were individually compared. Concordant data were flagged for merging in order to eliminate duplicate data (C) Mismatches were used to identify errors in the manual entry dataset. (D) We compared across all pathology fields for individual patients.

Analysis of Error within Pathology Reports

In order to gain insight into the error inherent within the original pathology reports from which we sourced data, we measured the concordance between pathological stage and one of the descriptors that lead to stage determination, namely the extraprostatic extension variable. An error was detected in cases of incongruity where extraprostatic extension was present but the staging was T2, or extraprostatic extension was absent but the staging was T3. We identified and excluded the small number of cases of stage T3b disease where seminal vesicle invasion was apparent but extraprostatic extension not definite, as this would not be considered an error. We examined only the electronically imported data read directly from the synoptic reports, so that the effect of manual data

entry was excluded. We identified a total of 971 cases where both pathological stage and extraprostatic extension fields had both been successfully imported from the synoptic pathology reports (Table I). Once again using Visual Basic, we generated a report indicating cases where there was mismatch between pathological stage and extraprostatic extension status. The original cases in which these mismatches occurred were all reviewed by a pathologist to confirm the presence of error in the source material.

Table I

Pathological Stage	**Number**
T2a	131
T2b	4
T2c	543
T3a	225
T3b	68
T4	0
Extraprostatic Extension	
Absent	670
Present	302
Total Cases for Comparison	971

Statistical Analysis

Percentage error rates were calculated by dividing the absolute number of errors by the total number of data points examined overall and in each field. Binomial distribution was used to calculate 95% confidence intervals for these rates, and Fisher's Exact Test applied to 2x2 contingency tables where necessary (PASW Statistics 18.0; IBM, Chicago, Illinois, 2010). All statistical tests were two-tailed and significance was assumed at $\alpha<0.05$.

RESULTS

Of the 421 patients selected for this part of the study, 320 had completely concordant data between manual entry and electronic import methods in a median of 12 pathology fields (range 10-13), indicating an outright accuracy in 76% of all patients. Seventy one patients (16.8%) had errors in one field only, whilst 18 (4.3%) had two or more incorrect fields, and 12 (2.9%) had 3-5 errors.

Analysis of error rate in each individual pathology field yielded rates of error ranging from 0.5-6.4%. Across all fields, the error rate was 2.8% (Table II). Assuming that errors in different fields occurred independently of one another, that the fraction of records where at least one error occurred would be given by $1\text{-}(1\text{-}p)^n$, where p is the overall error rate and n is the number of fields. In this case, $1\text{-}(1\text{-}0.028)^{12} = 31\%$. Since the proportion of records where error occurred was 24%, the errors appear not to occur independently across the fields.

Fields involving descriptive parameters appeared more error-prone than those with direct measurements or involving numerical figures, so we grouped the fields based on data format. Of the 2658 data points involving numbers (numeric and alphanumeric), 30 (1.1%, 95% CI 0.78-1.6) were erroneous, compared with 116 (4.7%, 95% CI 3.9-5.6) of the 2490 data points with text (p<0.0001). The five fields that required an element of interpretation in data entry also appeared more error-prone and again, when data was pooled, their difference in error rates compared with fields allowing for direct transcription was significantly greater (5.2% vs 1.3%, p<0.0001).

Table II

Pathology Field	Variable Type	Data Format	Total Data Points	Error	Error Rate (%)	(95% CI)
Gleason 1	Categorical	Numeric	415	2	0.5%	(0.06-1.7)
Gleason 2	Categorical	Numeric	415	3	0.7%	(0.15-2.1)
Gleason score	Categorical	Numeric	415	1	0.2%	(0.01-1.3)
Extraprostatic Extension*	Binary	Text	421	21	5.0%	(3.1-7.5)
Stage	Categorical	Alphanumeric	421	13	3.1%	(1.7-5.2)
Focality	Binary	Text	421	9	2.1%	(1.0-4.0)
Perineural Invasion*	Categorical	Text	421	27	6.4%	(4.3-9.2)
Lymphovascular Invasion*	Categorical	Text	421	27	6.4%	(4.3-9.2)
Prostatic Intraepithelial Neoplasia*	Categorical	Text	420	27	6.4%	(4.3-9.2)
Margins*	Binary	Text	386	5	1.3%	(0.42-3.0)
Tumour Volume	Continuous	Numeric	310	4	1.3%	(0.35-3.3)
Prostate Dimensions+	Continuous	Numeric	272	2	0.7%	(0.09-2.6)
Prostate Weight	Continuous	Numeric	410	5	1.2%	(0.40-2.8)
All Fields			5148	146	2.8%	(2.4-3.3)

** Data required some interpretation on data entry – these were coded numerically*
+Each data point was a combination of 3 numbers. Error never occurred in more than one dimension.

In the 971 cases used for the analysis of source data error, six cases were staged T2 but in fact were positive for extraprostatic extension (6 of 672, 0.9%). On pathologist review, these cases had indeed been understaged. One case of a T3 prostate cancer erroneously stated on the synoptic report that extraprostatic extension was not identified (Table III). This was the sole inconsistency between the original prose pathology report and its accompanying synoptic report (1 of 971, 0.1%). Although only two variables have been analysed, these figures suggest a very low rate of baseline error inherent in the pathology reports.

Table III

Pathological Stage	Matches	Mismatches	Error Rate %	(95% CI)
T2	672	6	0.9%	(0.33-1.9)
T3	292	1	0.3%	(0.01-1.9)
Total	964	7	0.7%	(0.30-1.5)

DISCUSSION

In a large contemporary radical prostatectomy dataset we have examined pathology data in a subset of over 400 patients and found the overall error rate due to manual data entry to be 2.8% across all fields. Individual fields were found to vary in error rates between 0.5% and 6.4%, and those involving descriptive text or requiring an element of interpretation appeared more vulnerable to error. Almost a quarter of patients had at least one data error when all pathology fields were considered, as might occur when multivariable statistical analysis is undertaken. We have also examined the source data electronically without human influence and established a baseline error rate of less than 1%.

The strengths of our study include a combination of factors that enable a realistic representation of a small to moderate sized oncology database used for research purposes. As the data was stored in a simple spreadsheet, not collected for clinical use and was sourced from primary clinical documents, this context of data entry represents a common scenario predating modern informatics solutions. We also examined a distinct set of data fields with varying formats important to clinical oncological research. Together, these increase the relevance of our findings to cancer datasets in general, and in particular to data which has provenance in times prior to the introduction of more sophisticated modes of data entry. In addition, we have checked the integrity of one aspect of our source data, which is of importance in both clinical and academic settings. This helps to set a lower limit to the general error rate that can be achieved with interventions for data integrity imposed beyond initial pathology reporting.

Our study was limited by its use of a single spreadsheet from a single series of patients. Although different institutions may use different data systems, the maintenance of clinical datasets on such spreadsheets is common in the clinical environment. Our source data was in the form of synoptic reports designed for ease of data transcription, rather than traditional pathology reports in prose and this may have reduced the true error rate of such data. Other major limitations were in the study design, whereby we could not differentiate easily between different types of data entry error despite being able to infer this to some extent from the format of data. Due to the nature of spreadsheets, we could not definitely account for row or column shifts in blocks of data as a source of error, although, on visual inspection of the errors this did not seem to be the case. As we only examined pathology fields covered by electronic import, the findings were not representative of the entire dataset, which also includes operative and perioperative details, and thus the study was not designed to test the effect that these other factors may have had on error nor was it designed to detect errors in these important areas. A final limitation was that the size of error in fields containing continuous data was not measured as we only identified mismatches in the datasets, and this is required to assess more fully any impact of error in those fields.

Studies of error in clinical datasets are scarce, owing in part to the time and resources required to conduct these audits. Our overall error rate in manually entered data appears similar to that of previous studies. In one study occurring over 10 years, Zellner et al reported an estimated probability of error in two systems at about 2.4% and the estimated error frequency in a database alone was 2.7%[15]. In this case, less than 10% of the overall dataset was examined using random sampling. Arndt et al performed a

detailed study of observer rating scores in a multicentre field setting[11], whilst Goldberg et al examined several clinical research databases, with errors ranging from 2.3% to 26.9% detected by the double-entry method in fields relating to timepoints of disease and tumour recurrence status[7]. In general, these studies have involved more sophisticated data entry interfaces that allowed more detailed analysis of the underlying aetiology of data errors.

In contrast, our study has directly examined most fields in the subset of pathology variables, of particular importance in oncology research, and removed the effects of manual transfer of data in the generation of the comparison dataset. We also analysed error in the source data, as they might exceed those of data entry and render attempts to decrease downstream error frequency less meaningful. In this case the rate of 0.9% in mismatch between stage and extraprostatic extension was reassuringly lower than the overall manual entry dataset error frequency, and was also lower than the generally cited rate of 1.4% error for prostate pathology[16].

Although an analysis of the impact of data errors on outcomes was an area our study was unable to address, as follow up times were too short in our dataset for meaningful results, others have demonstrated the variable effects that erroneous data might have on outcomes such as rates of tumour recurrence and mortality rates[7 8]. While it is likely that a low rate of data error will have little effect in univariable analysis, studies involving many fields and demonstrating a small effect size with borderline significance levels are intuitively liable to the effect of errors. In these cases, and particularly where accuracy across a large feature space is essential such as in translational genomics research, it is

preferable that data errors are accounted for. Some investigators have developed corrective statistical tools to be used with a specified error rate in source databases in a particular circumstance[8], but such tools are unlikely to become widely applicable without more reporting of error rates within different types of datasets and analyses of outcome differences.

The greatest influence on error rates in our own clinical dataset was the transition to electronic data feeds from clinical sources and the application of software to retrospectively replace manually entered data. In doing so, we decreased the portion of patients with manually entered data from 58% (853/1471) to 9% (128/1471), although for various technical reasons many fields still remain manually inputted amongst the earlier patients. With advances in technology, it may even be possible to extract data from even earlier pathology reports, since all reports are typewritten, and maintain a dataset with virtually no manually entered pathology data. Where such manually entered data is unavoidable or forms part of a larger dataset, due acknowledgement of the provenance of data from different parts of that dataset by performing separate analysis or by employing sensitivity analysis might be considered in research. In addition, the judicious selection of pathology fields based on liability to error might be used.

The recognition that direct use of clinically attained data leads to better accuracy is not new. The need to re-key data between clinical sources and database interfaces has long been acknowledged to be a significant source of human error[17], and the removal of this aspect of data entry would presumably increase accuracy of clinical datasets overall. In recent times, the availability of data directly collected in the clinical setting for other

healthcare activities including medical research has increased. One clinical group in a peripheral hospital centralised data collection for audit and research purposes via web-browser based application software and extensively integrated the system in daily use. In just 12 months, their unit amassed over 3000 near complete patient records and reported enhanced accuracy due to the demonstration of the immediate clinical value of high-quality data capture to the users[18]. Such integrated record systems have been shown to have additional clinical benefits[19], whilst data collected as near in time and space as possible to the point of care is known to improve overall accuracy[20]. Indeed, pathologists currently generate pathology reports prospectively as part of the clinical process, and our use of electronic data feeds from our service provider is an example of how clinical data can be directly captured for clinical or research use without the need for manual data entry.

With increasing drive for the widespread implementation of electronic health records[21], comes the opportunity for more electronic data feeds into data repositories from health services such as those used in the present study, and this effectively diminishes the reliance on manual data entry. However, as our study demonstrated, even a clinical data source itself has a pervasive error rate, and there will remain a need for active error trapping. Furthermore, retrospective replacement of manually entered data may afford opportunities to examine the errors in manually entered data. Perhaps this work would allow the development of tools to adequately account for errors in early datasets that no technology can correct.

CONCLUSION

We have evaluated a large radical prostatectomy dataset and while the overall rate of error was low, individual pathology fields were variably prone to error. We have demonstrated the feasibility of checking source clinical data for error, and the possibility of attaining high quality data using electronic data feeds for both prospective and retrospective parts of our data repository. We found that numerical data or data with fixed field entry provides better quality concordance between manual and electronic data-entry.

ACKNOWLEDGEMENTS

Dr M.K.Hong was supported by scholarships from the Royal Australasian College of Surgeons and from the National Health and Medical Research Council, Australia. We gratefully acknowledge Roberto Leiva for his technical assistance.

CONFLICT OF INTEREST STATEMENT

No conflict of interest to declare.

CONTRIBUTORSHIP

Matthew KH Hong

Conceived and designed the study, interpreted the data, wrote manuscript

Henry HI Yao

Study design, data interpretation, critically reviewed article, revised article, final approval

John S Pedersen

Study design, Acquired and analysed data, critically reviewed article, final approval

Justin S Peters Acquired data, critically reviewed article, final approval

Anthony J Costello

Acquired data, critically reviewed article, final approval

Declan G Murphy

Conception and design of study, critically reviewed article, final approval

Christopher M Hovens

Study design, data interpretation, revised article, final approval

Niall M Corcoran

Study design and conception, acquisition of data, interpretation of data, revised article, final approval

DATA SHARING

No additional data available.

FUNDING

None

References

1. Hudson TJ, Anderson W, Artez A, et al. International network of cancer genome projects. *Nature* 2010;464(7291):993-8.
2. Harel A, Dalah I, Pietrokovski S, et al. Omics data management and annotation. *Methods Mol Biol* 2011;719:71-96.
3. Needham DM, Sinopoli DJ, Dinglas VD, et al. Improving data quality control in quality improvement projects. *Int J Qual Health Care* 2009;21(2):145-50.
4. Arts DG, De Keizer NF, Scheffer GJ. Defining and improving data quality in medical registries: a literature review, case study, and generic framework. *J Am Med Inform Assoc* 2002;9(6):600-11.
5. Beretta L, Aldrovandi V, Grandi E, et al. Improving the quality of data entry in a low-budget head injury database. *Acta Neurochir (Wien)* 2007;149(9):903-9.
6. Seddon DJ, Williams EM. Data quality in population-based cancer registration: an assessment of the Merseyside and Cheshire Cancer Registry. *Br J Cancer* 1997;76(5):667-74.
7. Goldberg SI, Niemierko A, Turchin A. Analysis of data errors in clinical research databases. *AMIA Annu Symp Proc* 2008:242-6.
8. Gallivan S, Pagel C. Modelling of errors in databases. *Health Care Management Science* 2008;11(1):35-40.
9. Goldberg SI, Niemierko A, Shubina M, et al. "Summary Page": a novel tool that reduces omitted data in research databases. *BMC Med Res Methodol* 2010;10:91.
10. Warsi AA, White S, McCulloch P. Completeness of data entry in three cancer surgery databases. *Eur J Surg Oncol* 2002;28(8):850-6.
11. Arndt S, Tyrrell G, Woolson RF, et al. Effects of errors in a multicenter medical study: preventing misinterpreted data. *J Psychiatr Res* 1994;28(5):447-59.
12. Hayrinen K, Saranto K, Nykanen P. Definition, structure, content, use and impacts of electronic health records: a review of the research literature. *Int J Med Inform* 2008;77(5):291-304.
13. Antonarakis ES, Feng Z, Trock BJ, et al. The natural history of metastatic progression in men with prostate-specific antigen recurrence after radical prostatectomy: long-term follow-up. *BJU Int* 2012;109(1):32-9.
14. Neo EL, Beeke C, Price T, et al. South Australian clinical registry for metastatic colorectal cancer. *ANZ J Surg* 2011;81(5):352-7.
15. Zellner D, Schromm T, Frankewitsch T, et al. Structured data entry for reliable acquisition of pharmacokinetic data. *Methods Inf Med* 1996;35(3):261-4.
16. Frable WJ. Surgical pathology--second reviews, institutional reviews, audits, and correlations: what's out there? Error or diagnostic variation? *Arch Pathol Lab Med* 2006;130(5):620-5.
17. Gostel R. HyperCard to SPSS: improving data integrity. *Comput Nurs* 1993;11(1):25-8.
18. Tran P, Morrison SG, Lade JA, et al. An integrated approach to surgical audit. *ANZ J Surg* 2011;81(5):313-4.
19. Featherstone I, Keen J. Do integrated record systems lead to integrated services? An observational study of a multi-professional system in a diabetes service. *Int J Med Inform* 2012;81(1):45-52.

20. Veen EJ, Janssen-Heijnen ML, Leenen LP, et al. The registration of complications in surgery: a learning curve. *World J Surg* 2005;29(3):402-9.

21. Pearce C, Haikerwal MC. E-health in Australia: time to plunge into the 21st century. *Med J Aust* 2010;193(7):397-8.

Error Rates in a Clinical Data Repository: Lessons from the Transition to Electronic Data Transfer

Matthew KH Hong
Henry HI Yao
John S Pedersen[*]
Justin S Peters
Anthony J Costello
Declan G Murphy[+]
Christopher M Hovens
Niall M Corcoran

Division of Urology, Department of Surgery, University of Melbourne, Royal Melbourne Hospital and the Australian Prostate Cancer Research Centre Epworth, Victoria, Australia.
* TissuPath Specialist Pathology, Mount Waverley and Monash University Faculty of Medicine, Victoria, Australia
+ Peter MacCallum Cancer Centre, East Melbourne, Australia

Correspondence:
Dr Matthew Hong,
Department of Surgery
University of Melbourne
Royal Melbourne Hospital
Grattan St Parkville
VIC 3050 Australia
T: 613 9342 7703
E: m.k.hong@ausdoctors.net

Keywords: Database, prostate cancer, data quality, error sources, clinical informatics
Word count: 3927 words

ABSTRACT

Objective: Data errors are a well-documented part of clinical datasets as is their potential to confound downstream analysis. In this study we explore the reliability of manually transcribed data across different pathology fields in a prostate cancer database and also measure error rates attributable to the source data.

Design: Descriptive study

Setting: Specialist urology service at a single centre in metropolitan Victoria in Australia

Participants: Between 2004 and 2011, 1471 patients underwent radical prostatectomy at our institution. In a large proportion of these cases, clinicopathological variables were recorded by manual data-entry. In 2011, we obtained electronic versions of the same printed pathology reports for our cohort. The data were electronically imported in parallel to any existing manual entry record enabling direct comparison between them.

Outcome measures: Error rates of manually entered data compared with electronically imported data across clinicopathological fields.

Results: 421 patients had at least 10 comparable pathology fields between the electronic import and manual records and were selected for study. 320 patients had concordant data between manually entered and electronically populated fields in a median of 12 pathology fields (range 10-13), indicating an outright accuracy in manually entered pathology data in 76% of patients. Across all fields, the error rate was 2.8% whilst individual field error ranges from 0.5-6.4%. Fields in text formats were significantly more error-prone than those with direct measurements or involving numerical figures ($p<0.001$). 971 cases were available for review of error within the source data, with figures of 0.1%-0.9%.

Conclusion: While the overall rate of error was low in manually entered data, individual pathology fields were variably prone to error. High quality pathology data can be obtained for both prospective and retrospective parts of our data repository and the electronic checking of source pathology data for error is feasible.

BACKGROUND AND SIGNIFICANCE

The majority of clinical research publications are based on the analysis of prospectively collected, clinical databases. In addition, patient centred databases are increasingly important in translational research efforts, as appropriately annotated tissue banks are the foundation for global multi-institutional collaborative efforts in genetic and epigenetic screening of various diseases[1]. Yet despite the stringent quality controls placed on the vast amounts of research data derived from these studies and the acute awareness of the need to control data quality[2 3], the inherent accuracy of original clinical datasets is one area that receives relatively little attention.

Data errors are common in clinical datasets[4-6], with some cancer databases recording error rates as high as almost 27% in some fields [7]. Such errors have the potential to adversely affect data analysis and interpretation, and can lead to erroneous conclusions[8]. Methods to first identify then correct errors in these datasets would be immensely valuable in the setting of the large-scale genomics projects being performed.

Two types of errors are described in the literature: one of omission, and one of erroneous value. Although it is sometimes argued that missing values carry greater impact due to their greater prevalence[9], which may be up to 55% in cancer surgery databases[10], these errors are more easily detected with judicious computer queries and corrected with retrospective data collection. On the contrary, once erroneous values permeate a dataset, their effects can cascade in unpredictable ways. Errors in high impact fields have been shown to adversely affect the interpretation of statistical analyses, even if the errors are at low prevalence[11]. Whilst it is well known that

structured data entry improves the accuracy of manual documentation[12], much of the clinical data of high value to researchers predates any effective informatics solutions aimed at data quality that might exist today. Instead, manual retrospective transcription of data from clinical records into relatively unstructured spreadsheets constitutes the data entry method for many clinical audits that subsequently serve research purposes. These datasets may have even transitioned to more carefully constructed data entry interfaces, as might occur in conditions such as prostate cancer where long follow up times of over ten years are necessary for study of oncological outcomes[13]. In such cases, the provenance of the data collected with earlier means may not be accounted for with subsequent analysis.

Studies involving large cancer datasets rarely report error rates or their management, and it is difficult to assess the impact that these may have on the outcomes reported[14]. Given the considerable effort that generally goes towards the collection of data for a large clinical database, it is unsurprising that surplus resources are usually unavailable to place towards the check of data accuracy. Although larger numbers in databases may be used to counter the problem of errors, the combination of datasets, particularly with different fields would serve only to magnify error rates.

Knowledge of errors in manually collected data could give insight into how these may be accounted for in subsequent analysis. In cancer databases, pathology data are of particular importance as they are relied upon to build cohorts of clinical relevance in research. Often there are multiple fields that give equivalent indications of underlying biology and any one could be analysed to similar effect. For example, both the

percentage involvement of tumour or its measured size may serve as parameters of cancer burden. It is unlikely that different types of data would have equivalent vulnerability to error, and knowledge of the fields or types of fields that might be more error-prone with manual data entry could help researchers judiciously select fields for analysis based on greater accuracy. It might also aid informaticians to focus on error prevention in fields that carry particular importance in clinical and research settings. In addition, it is important to measure the baseline level of error inherent within the pathology reports themselves, the data source, as no degree of accuracy in manual transcription or even automated processes can result in a lower error rate without amendment of the original report.

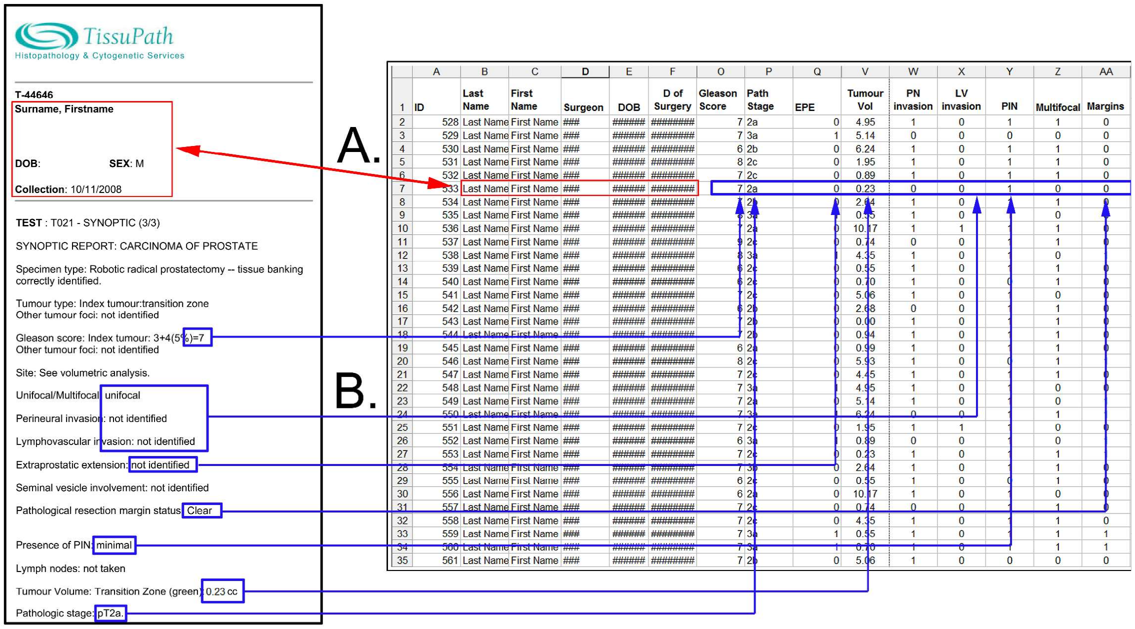
In this article we explored the reliability of manually transcribed pathology data across different fields in a large contemporary prostate cancer database. Initially housed as a Microsoft Excel spreadsheet, the database has evolved to become server based with a web-based interface. We have established automated electronic datafeeds from our pathology service provider to reduce the manual human data entry component in the pathology data, and we have used these to prospectively and retrospectively populate data fields. We compared overlapping data from the electronic feeds and previously manually entered data, whereby we could gauge the accuracy of pathology details in a subset of our patients. In this way, we could determine the error involved in manual data entry in different fields across patients with relative ease. In addition, we explored error that might be attributable to the source data.

PATIENT AND METHODS

Database Systems and Data Linking

Between 2004 and 2011, 1471 patients underwent radical prostatectomy across our institutions. Variables including demographics, pre-operative PSA, pathological Gleason score and stage, and other pathological data relating to the prostatectomy specimens were manually entered in a non-relational database (Microsoft Excel spreadsheet) for the first 853 of these cases with 57 fields per patient (2004-2008). Although most data collection was primarily prospectively performed, pathology data was obtained retrospectively once printed specimen pathology reports became available, or missing data was found on later review. For every patient, printed pathology reports were consulted and data manually transcribed into the spreadsheet. Each of the reports were issued by a single pathology group and consisted of one to two pages of prose. Since 2006, the reports have been accompanied by a separate page with a 'synoptic' report. This synoptic report contained the pathology data in a structured format with fields of interest listed on a single page enabling greater ease of interpretation over the traditional reports in prose. Manual data entry was performed by four surgical residents with knowledge of prostate cancer pathology and versed in the relevant terminology. ~~Several different junior medical staff performed manual data entry at sporadic times throughout.~~ In 2010 our institution moved all data to Caisis 5.0, a web-based relational database system developed at the Memorial Sloan Kettering Cancer Centre in New York, and ceased manual recording of pathology data from hardcopy reports.

We subsequently established a data link between our database and the pathology group whereby electronically encrypted reports were provided in HL7 standard v2.31 format, a health industry information technology standard. The reports were retrieved using client/server software through a TCP/IP link. Custom software was developed in Visual Basic (Microsoft Visual Studio 2010) that enabled us to parse text or values of interest from the synoptic reports and automatically populate associated fields in our database system. (**Fig 1**). All new pathology data henceforth was imported in this way after testing of the new software for accuracy was performed. Digital import of data was conservative, and the software was written in such a way that data was not transferred in cases of ambiguity. Given that the original synoptic reports dating back to 2006 were digitally stored by our pathology group, we were able to import pathology data from this time period with our software and use it in preference to any data arising from manual data entry for patients from 2006 to 2011.



TissuPath
Histopathology & Cytogenetic Services

T-44646
Surname, Firstname

DOB: **SEX**: M

Collection: 10/11/2008

TEST : T021 - SYNOPTIC (3/3)

SYNOPTIC REPORT: CARCINOMA OF PROSTATE

Specimen type: Robotic radical prostatectomy -- tissue banking correctly identified.

Tumour type: Index tumour:transition zone
Other tumour foci: not identified

Gleason score: Index tumour: 3+4(5%)=7
Other tumour foci: not identified

Site: See volumetric analysis.

Unifocal/Multifocal: unifocal

Perineural invasion: not identified

Lymphovascular invasion: not identified

Extraprostatic extension: not identified

Seminal vesicle involvement: not identified

Pathological resection margin status: Clear

Presence of PIN: minimal

Lymph nodes: not taken

Tumour Volume: Transition Zone (green): 0.23 cc

Pathologic stage: pT2a.

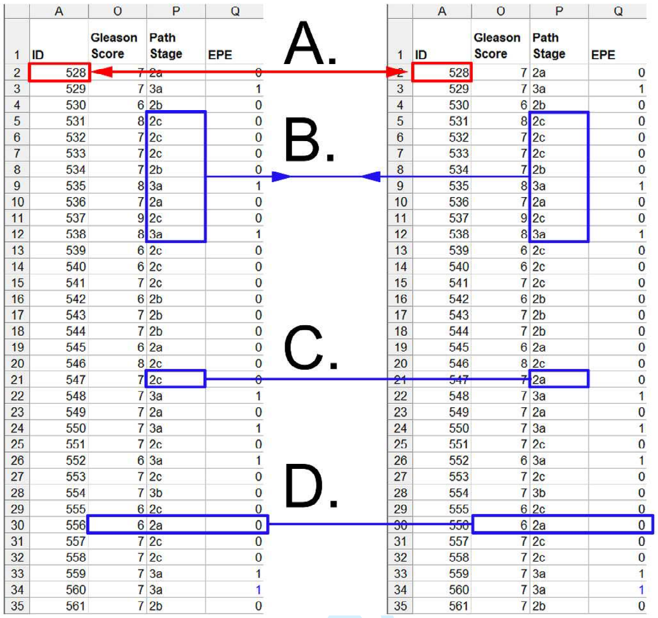
	A	B	C	D	E	F	O	P	Q	V	W	X	Y	Z	AA
1	ID	Last Name	First Name	Surgeon	DOB	D of Surgery	Gleason Score	Path Stage	EPE	Tumour Vol	PN invasion	LV invasion	PIN	Multifocal	Margins
2	528	Last Name	First Name	###	######	########	7	2a	0	4.95	1	0	1	1	0
3	529	Last Name	First Name	###	######	########	7	3a	1	5.14	0	0	0	0	0
4	530	Last Name	First Name	###	######	########	6	2b	0	6.24	1	0	1	1	0
5	531	Last Name	First Name	###	######	########	8	2c	0	1.95	1	0	1	1	0
6	532	Last Name	First Name	###	######	########	7	2c	0	0.89	1	0	1	1	0
7	533	Last Name	First Name	###	######	########	7	2a	0	0.23	0	0	1	0	0
8	534	Last Name	First Name	###	######	########	7	2b	0	2.64	1	0	1	1	0
9	535	Last Name	First Name	###	######	########	8	3a	1	0.55	1	0	1	0	[illegible]
10	536	Last Name	First Name	###	######	########	7	2a	0	10.17	1	1	1	1	0
11	537	Last Name	First Name	###	######	########	9	2c	0	0.74	0	0	1	1	0
12	538	Last Name	First Name	###	######	########	8	3a	1	4.35	1	0	1	0	1
13	539	Last Name	First Name	###	######	########	6	2c	0	0.55	1	0	1	1	0
14	540	Last Name	First Name	###	######	########	6	2c	0	0.70	1	0	0	1	0
15	541	Last Name	First Name	###	######	########	7	2c	0	5.06	1	0	1	0	0
16	542	Last Name	First Name	###	######	########	6	2b	0	2.68	0	0	1	1	0
17	543	Last Name	First Name	###	######	########	7	2b	0	0.00	1	0	1	1	0
18	544	Last Name	First Name	###	######	########	7	2b	0	0.94	1	0	1	1	0
19	545	Last Name	First Name	###	######	########	6	2a	0	0.99	1	0	1	1	0
20	546	Last Name	First Name	###	######	########	8	2c	0	5.93	1	0	0	1	1
21	547	Last Name	First Name	###	######	########	7	2c	0	4.45	1	0	1	1	0
22	548	Last Name	First Name	###	######	########	7	3a	1	4.95	1	0	1	0	0
23	549	Last Name	First Name	###	######	########	7	2a	0	5.14	1	0	1	0	1
24	550	Last Name	First Name	###	######	########	7	3a	1	6.24	0	0	1	1	1
25	551	Last Name	First Name	###	######	########	7	2c	0	1.95	1	1	1	0	0
26	552	Last Name	First Name	###	######	########	6	3a	1	0.89	0	0	1	0	1
27	553	Last Name	First Name	###	######	########	7	2c	0	0.23	1	0	1	1	1
28	554	Last Name	First Name	###	######	########	7	3b	0	2.64	1	0	1	1	0
29	555	Last Name	First Name	###	######	########	6	2c	0	0.55	1	0	0	1	0
30	556	Last Name	First Name	###	######	########	6	2a	0	10.17	1	0	1	0	0
31	557	Last Name	First Name	###	######	########	7	2c	0	0.74	0	0	1	1	0
32	558	Last Name	First Name	###	######	########	7	2c	0	4.35	1	0	1	1	0
33	559	Last Name	First Name	###	######	########	7	3a	1	0.55	1	0	1	1	1
34	560	Last Name	First Name	###	######	########	7	3a	1	0.70	1	0	1	1	1
35	561	Last Name	First Name	###	######	########	7	2b	0	5.06	1	0	0	0	0

Fig. 1 Schematic representation of the digital import of pathology data. Structured ‘synoptic’ reports facilitated digital recognition of relevant pathology fields. (A) Demographics data was used to link reports to individual patients in the database and

(B) individual data were then extracted from the report and directed to populate relevant fields in the main database.

Analysis of Error in Manual Data Entry

The digitally imported data were placed parallel to any existing manual entry records and following institutional review board approval, we were able to directly compare between digitally imported data and manual entry data for 752 patients where records from the two data entry methods co-existed. We assumed that any mismatch in the fields within the manual entry and digital import was due to human error in the data entry, as data were copied from printed versions of these same reports in the first instance. The importing software had been extensively tested and errors would be systematic within each field rather than transcriptional in nature. We excluded from analysis specimen pathology fields with fewer than 200 comparable entries in order to detect at least a 0.5% error rate. 421 patients had at least 10 completed pathology fields in both manual entry and electronic import records and were thus selected for study. This would allow us to detect the error rate across a patient's fields and also minimise individual patient factors in explaining error rates in different fields. Within each pathology field, we linked records based on unique identifiers and electronically compared them using custom prepared software (Visual Basic, Microsoft Visual Studio 2010). We identified and counted any mismatches and then compared across all fields for each patient to determine the number of patients affected by one or more errors across the cohort (**Fig. 2**).



	A	O	P	Q
1	ID	Gleason Score	Path Stage	EPE
2	528	7	2a	0
3	529	7	3a	1
4	530	6	2b	0
5	531	8	2c	0
6	532	7	2c	0
7	533	7	2c	0
8	534	7	2b	0
9	535	8	3a	1
10	536	7	2a	0
11	537	9	2c	0
12	538	8	3a	1
13	539	6	2c	0
14	540	6	2c	0
15	541	7	2c	0
16	542	6	2b	0
17	543	7	2b	0
18	544	7	2b	0
19	545	6	2a	0
20	546	8	2c	0
21	547	7	2c	0
22	548	7	3a	1
23	549	7	2a	0
24	550	7	3a	1
25	551	7	2c	0
26	552	6	3a	1
27	553	7	2c	0
28	554	7	3b	0
29	555	6	2c	0
30	556	6	2a	0
31	557	7	2c	0
32	558	7	2c	0
33	559	7	3a	1
34	560	7	3a	1
35	561	7	2b	0

	A	O	P	Q
1	ID	Gleason Score	Path Stage	EPE
2	528	7	2a	0
3	529	7	3a	1
4	530	6	2b	0
5	531	8	2c	0
6	532	7	2c	0
7	533	7	2c	0
8	534	7	2b	0
9	535	8	3a	1
10	536	7	2a	0
11	537	9	2c	0
12	538	8	3a	1
13	539	6	2c	0
14	540	6	2c	0
15	541	7	2c	0
16	542	6	2b	0
17	543	7	2b	0
18	544	7	2b	0
19	545	6	2a	0
20	546	8	2c	0
21	547	7	2a	0
22	548	7	3a	1
23	549	7	2a	0
24	550	7	3a	1
25	551	7	2c	0
26	552	6	3a	1
27	553	7	2c	0
28	554	7	3b	0
29	555	6	2c	0
30	556	6	2a	0
31	557	7	2c	0
32	558	7	2c	0
33	559	7	3a	1
34	560	7	3a	1
35	561	7	2b	0

Fig. 2 Schematic representation of the comparison of a dataset imported digitally and in parallel to a manually entered dataset. (A) Records were linked using unique patient identifiers, and (B) pathology fields were individually compared. Concordant data were flagged for merging in order to eliminate duplicate data (C) Mismatches were used to identify errors in the manual entry dataset. (D) We compared across all pathology fields for individual patients.

Analysis of Error within Pathology Reports

In order to gain insight into the error inherent within the original pathology reports from which we sourced data, we measured the concordance between pathological stage and one of the descriptors that lead to stage determination, namely the extraprostatic extension variable. An error was detected in cases of incongruity where extraprostatic extension was present but the staging was T2, or extraprostatic extension was absent but the staging was T3. We identified and excluded the small number of cases of stage T3b disease where seminal vesicle invasion was apparent but extraprostatic extension not definite, as this would not be considered an error. We examined only the electronically imported data read directly from the synoptic reports, so that the effect of manual data

entry was excluded. We identified a total of 971 cases where both pathological stage and extraprostatic extension fields had both been successfully imported from the synoptic pathology reports (Table I). Once again using Visual Basic, we generated a report indicating cases where there was mismatch between pathological stage and extraprostatic extension status. The original cases in which these mismatches occurred were all reviewed by a pathologist to confirm the presence of error in the source material.

Table I

Pathological Stage	**Number**
T2a	131
T2b	4
T2c	543
T3a	225
T3b	68
T4	0
Extraprostatic Extension	
Absent	670
Present	302
Total Cases for Comparison	971

Statistical Analysis

Percentage error rates were calculated by dividing the absolute number of errors by the total number of data points examined overall and in each field. Binomial distribution was used to calculate 95% confidence intervals for these rates, and Fisher's Exact Test applied to 2x2 contingency tables where necessary (PASW Statistics 18.0; IBM, Chicago, Illinois, 2010). All statistical tests were two-tailed and significance was assumed at $\alpha<0.05$.

RESULTS

Of the 421 patients selected for this part of the study, 320 had completely concordant data between manual entry and electronic import methods in a median of 12 pathology fields (range 10-13), indicating an outright accuracy in 76% of all patients. Seventy one patients (16.8%) had errors in one field only, whilst 18 (4.3%) had two or more incorrect fields, and 12 (2.9%) had 3-5 errors.

Analysis of error rate in each individual pathology field yielded rates of error ranging from 0.5-6.4%. Across all fields, the error rate was 2.8% (Table II). Assuming that errors in different fields occurred independently of one another, that the fraction of records where at least one error occurred would be given by $1\text{-}(1\text{-}p)^n$, where p is the overall error rate and n is the number of fields. In this case, $1\text{-}(1\text{-}0.028)^{12} = 31\%$. Since the proportion of records where error occurred was 24%, the errors appear not to occur independently across the fields.

Fields involving descriptive parameters appeared more error-prone than those with direct measurements or involving numerical figures, so we grouped the fields based on data format. Of the 2658 data points involving numbers (numeric and alphanumeric), 30 (1.1%, 95% CI 0.78-1.6) were erroneous, compared with 116 (4.7%, 95% CI 3.9-5.6) of the 2490 data points with text (p<0.0001). The five fields that required an element of interpretation in data entry also appeared more error-prone and again, when data was pooled, their difference in error rates compared with fields allowing for direct transcription was significantly greater (5.2% vs 1.3%, p<0.0001).

Table II

Pathology Field	Variable Type	Data Format	Total Data Points	Error	Error Rate (%)	(95% CI)
Gleason 1	Categorical	Numeric	415	2	0.5%	(0.06-1.7)
Gleason 2	Categorical	Numeric	415	3	0.7%	(0.15-2.1)
Gleason score	Categorical	Numeric	415	1	0.2%	(0.01-1.3)
Extraprostatic Extension*	Binary	Text	421	21	5.0%	(3.1-7.5)
Stage	Categorical	Alphanumeric	421	13	3.1%	(1.7-5.2)
Focality	Binary	Text	421	9	2.1%	(1.0-4.0)
Perineural Invasion*	Categorical	Text	421	27	6.4%	(4.3-9.2)
Lymphovascular Invasion*	Categorical	Text	421	27	6.4%	(4.3-9.2)
Prostatic Intraepithelial Neoplasia*	Categorical	Text	420	27	6.4%	(4.3-9.2)
Margins*	Binary	Text	386	5	1.3%	(0.42-3.0)
Tumour Volume	Continuous	Numeric	310	4	1.3%	(0.35-3.3)
Prostate Dimensions+	Continuous	Numeric	272	2	0.7%	(0.09-2.6)
Prostate Weight	Continuous	Numeric	410	5	1.2%	(0.40-2.8)
All Fields			5148	146	2.8%	(2.4-3.3)

** Data required some interpretation on data entry – these were coded numerically*
+Each data point was a combination of 3 numbers. Error never occurred in more than one dimension.

In the 971 cases used for the analysis of source data error, six cases were staged T2 but in fact were positive for extraprostatic extension (6 of 672, 0.9%). On pathologist review, these cases had indeed been understaged. One case of a T3 prostate cancer erroneously stated on the synoptic report that extraprostatic extension was not identified (Table III). This was the sole inconsistency between the original prose pathology report and its accompanying synoptic report (1 of 971, 0.1%). Although only two variables have been analysed, these figures suggest a very low rate of baseline error inherent in the pathology reports.

Table III

Pathological Stage	Matches	Mismatches	Error Rate %	(95% CI)
T2	672	6	0.9%	(0.33-1.9)
T3	292	1	0.3%	(0.01-1.9)
Total	964	7	0.7%	(0.30-1.5)

DISCUSSION

In a large contemporary radical prostatectomy dataset we have examined pathology data in a subset of over 400 patients and found the overall error rate due to manual data entry to be 2.8% across all fields. Individual fields were found to vary in error rates between 0.5% and 6.4%, and those involving descriptive text or requiring an element of interpretation appeared more vulnerable to error. Almost a quarter of patients had at least one data error when all pathology fields were considered, as might occur when multivariable statistical analysis is undertaken. We have also examined the source data electronically without human influence and established a baseline error rate of less than 1%.

The strengths of our study include a combination of factors that enable a realistic representation of a small to moderate sized oncology database used for research purposes. As the data was stored in a simple spreadsheet, not collected for clinical use and was sourced from primary clinical documents, this context of data entry represents a common scenario predating modern informatics solutions. We also examined a distinct set of data fields with varying formats important to clinical oncological research. Together, these increase the relevance of our findings to cancer datasets in general, and in particular to data which has provenance in times prior to the introduction of more sophisticated modes of data entry. In addition, we have checked the integrity of one aspect of our source data, which is of importance in both clinical and academic settings. This helps to set a lower limit to the general error rate that can be achieved with interventions for data integrity imposed beyond initial pathology reporting.

Our study was limited by its use of a single spreadsheet from a single series of patients. Although different institutions may use different data systems, the maintenance of clinical datasets on such spreadsheets is common in the clinical environment. Our source data was in the form of synoptic reports designed for ease of data transcription, rather than traditional pathology reports in prose and this may have reduced the true error rate of such data. Other major limitations were in the study design, whereby we could not differentiate easily between different types of data entry error despite being able to infer this to some extent from the format of data. Due to the nature of spreadsheets, we could not definitely account for row or column shifts in blocks of data as a source of error, although, on visual inspection of the errors this did not seem to be the case. As we only examined pathology fields covered by electronic import, the findings were not representative of the entire dataset, which also includes operative and perioperative details, and thus the study was not designed to test the effect that these other factors may have had on error nor was it designed to detect errors in these important areas. A final limitation was that the size of error in fields containing continuous data was not measured as we only identified mismatches in the datasets, and this is required to assess more fully any impact of error in those fields.

Studies of error in clinical datasets are scarce, owing in part to the time and resources required to conduct these audits. Our overall error rate in manually entered data appears similar to that of previous studies. In one study occurring over 10 years, Zellner et al reported an estimated probability of error in two systems at about 2.4% and the estimated error frequency in a database alone was 2.7%[15]. In this case, less than 10% of the overall dataset was examined using random sampling. Arndt et al performed a

detailed study of observer rating scores in a multicentre field setting[11], whilst Goldberg et al examined several clinical research databases, with errors ranging from 2.3% to 26.9% detected by the double-entry method in fields relating to timepoints of disease and tumour recurrence status[7]. In general, these studies have involved more sophisticated data entry interfaces that allowed more detailed analysis of the underlying aetiology of data errors.

In contrast, our study has directly examined most fields in the subset of pathology variables, of particular importance in oncology research, and removed the effects of manual transfer of data in the generation of the comparison dataset. We also analysed error in the source data, as they might exceed those of data entry and render attempts to decrease downstream error frequency less meaningful. In this case the rate of 0.9% in mismatch between stage and extraprostatic extension was reassuringly lower than the overall manual entry dataset error frequency, and was also lower than the generally cited rate of 1.4% error for prostate pathology[16].

Although an analysis of the impact of data errors on outcomes was an area our study was unable to address, as follow up times were too short in our dataset for meaningful results, others have demonstrated the variable effects that erroneous data might have on outcomes such as rates of tumour recurrence and mortality rates[7 8]. While it is likely that a low rate of data error will have little effect in univariable analysis, studies involving many fields and demonstrating a small effect size with borderline significance levels are intuitively liable to the effect of errors. In these cases, and particularly where accuracy across a large feature space is essential such as in translational genomics research, it is

preferable that data errors are accounted for. Some investigators have developed corrective statistical tools to be used with a specified error rate in source databases in a particular circumstance[8], but such tools are unlikely to become widely applicable without more reporting of error rates within different types of datasets and analyses of outcome differences.

The greatest influence on error rates in our own clinical dataset was the transition to electronic data feeds from clinical sources and the application of software to retrospectively replace manually entered data. In doing so, we decreased the portion of patients with manually entered data from 58% (853/1471) to 9% (128/1471), although for various technical reasons many fields still remain manually inputted amongst the earlier patients. With advances in technology, it may even be possible to extract data from even earlier pathology reports, since all reports are typewritten, and maintain a dataset with virtually no manually entered pathology data. Where such manually entered data is unavoidable or forms part of a larger dataset, due acknowledgement of the provenance of data from different parts of that dataset by performing separate analysis or by employing sensitivity analysis might be considered in research. In addition, the judicious selection of pathology fields based on liability to error might be used.

The recognition that direct use of clinically attained data leads to better accuracy is not new. The need to re-key data between clinical sources and database interfaces has long been acknowledged to be a significant source of human error[17], and the removal of this aspect of data entry would presumably increase accuracy of clinical datasets overall. In recent times, the availability of data directly collected in the clinical setting for other

healthcare activities including medical research has increased. One clinical group in a peripheral hospital centralised data collection for audit and research purposes via web-browser based application software and extensively integrated the system in daily use. In just 12 months, their unit amassed over 3000 near complete patient records and reported enhanced accuracy due to the demonstration of the immediate clinical value of high-quality data capture to the users[18]. Such integrated record systems have been shown to have additional clinical benefits[19], whilst data collected as near in time and space as possible to the point of care is known to improve overall accuracy[20]. Indeed, pathologists currently generate pathology reports prospectively as part of the clinical process, and our use of electronic data feeds from our service provider is an example of how clinical data can be directly captured for clinical or research use without the need for manual data entry.

With increasing drive for the widespread implementation of electronic health records[21], comes the opportunity for more electronic data feeds into data repositories from health services such as those used in the present study, and this effectively diminishes the reliance on manual data entry. However, as our study demonstrated, even a clinical data source itself has a pervasive error rate, and there will remain a need for active error trapping. Furthermore, retrospective replacement of manually entered data may afford opportunities to examine the errors in manually entered data. Perhaps this work would allow the development of tools to adequately account for errors in early datasets that no technology can correct.

CONCLUSION

We have evaluated a large radical prostatectomy dataset and while the overall rate of error was low, individual pathology fields were variably prone to error. We have demonstrated the feasibility of checking source clinical data for error, and the possibility of attaining high quality data using electronic data feeds for both prospective and retrospective parts of our data repository. We found that numerical data or data with fixed field entry provides better quality concordance between manual and electronic data-entry.

ACKNOWLEDGEMENTS

Dr M.K.Hong was supported by scholarships from the Royal Australasian College of Surgeons and from the National Health and Medical Research Council, Australia. We gratefully acknowledge Roberto Leiva for his technical assistance.

Article Summary

Article focus

- Although use of structured electronic databases is widespread, a substantial amount of clinical data used in research predates this.
- There is a paucity of literature on error rates in such clinical datasets used in research.
- We explored the reliability of manually transcribed data across different pathology fields in a prostate cancer database and also measured error rates attributable to the source data.

Key messages

- Whilst overall rate of error for manually entered data can be low, individual fields may be variably prone to error, especially those involving descriptive text or requiring an element of interpretation.
- Computerised systems can be used to check clinical source data for error.
- The use of electronic data feeds retrospectively can replace manually collected data fields in some cases to improve overall accuracy.

Strengths and limitations of this study

- Our study design provides a realistic representation of a small to moderate sized oncology database used for research purposes.
- We checked the integrity of one aspect of our source data.
- Our study was limited by its use of a single spreadsheet from a single series of patients.
- As we only examined pathology fields covered by electronic import, the findings were not representative of the entire dataset.

CONFLICT OF INTEREST STATEMENT

No conflict of interest to declare.

References

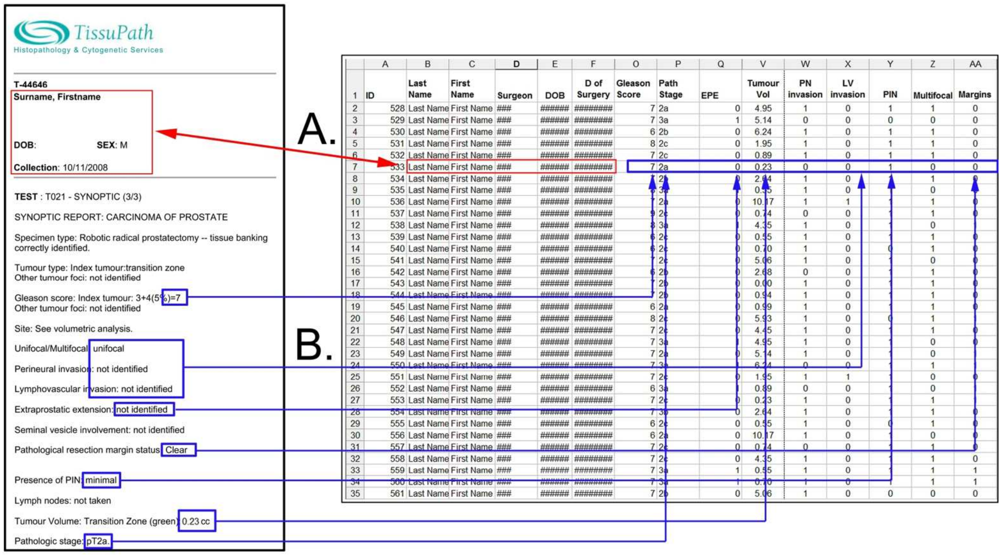
1. Hudson TJ, Anderson W, Artez A, Barker AD, Bell C, Bernabe RR, et al. International network of cancer genome projects. *Nature* 2010;464(7291):993-8.
2. Harel A, Dalah I, Pietrokovski S, Safran M, Lancet D. Omics data management and annotation. *Methods Mol Biol* 2011;719:71-96.
3. Needham DM, Sinopoli DJ, Dinglas VD, Berenholtz SM, Korupolu R, Watson SR, et al. Improving data quality control in quality improvement projects. *Int J Qual Health Care* 2009;21(2):145-50.
4. Arts DG, De Keizer NF, Scheffer GJ. Defining and improving data quality in medical registries: a literature review, case study, and generic framework. *J Am Med Inform Assoc* 2002;9(6):600-11.
5. Beretta L, Aldrovandi V, Grandi E, Citerio G, Stocchetti N. Improving the quality of data entry in a low-budget head injury database. *Acta Neurochir (Wien)* 2007;149(9):903-9.
6. Seddon DJ, Williams EM. Data quality in population-based cancer registration: an assessment of the Merseyside and Cheshire Cancer Registry. *Br J Cancer* 1997;76(5):667-74.
7. Goldberg SI, Niemierko A, Turchin A. Analysis of data errors in clinical research databases. *AMIA Annu Symp Proc* 2008:242-6.
8. Gallivan S, Pagel C. Modelling of errors in databases. *Health Care Management Science* 2008;11(1):35-40.
9. Goldberg SI, Niemierko A, Shubina M, Turchin A. "Summary Page": a novel tool that reduces omitted data in research databases. *BMC Med Res Methodol* 2010;10:91.
10. Warsi AA, White S, McCulloch P. Completeness of data entry in three cancer surgery databases. *Eur J Surg Oncol* 2002;28(8):850-6.
11. Arndt S, Tyrrell G, Woolson RF, Flaum M, Andreasen NC. Effects of errors in a multicenter medical study: preventing misinterpreted data. *J Psychiatr Res* 1994;28(5):447-59.
12. Hayrinen K, Saranto K, Nykanen P. Definition, structure, content, use and impacts of electronic health records: a review of the research literature. *Int J Med Inform* 2008;77(5):291-304.
13. Antonarakis ES, Feng Z, Trock BJ, Humphreys EB, Carducci MA, Partin AW, et al. The natural history of metastatic progression in men with prostate-specific antigen recurrence after radical prostatectomy: long-term follow-up. *BJU Int* 2012;109(1):32-9.
14. Neo EL, Beeke C, Price T, Maddern G, Karapetis C, Luke C, et al. South Australian clinical registry for metastatic colorectal cancer. *ANZ J Surg* 2011;81(5):352-7.
15. Zellner D, Schromm T, Frankewitsch T, Giehl M, Keller F. Structured data entry for reliable acquisition of pharmacokinetic data. *Methods Inf Med* 1996;35(3):261-4.
16. Frable WJ. Surgical pathology--second reviews, institutional reviews, audits, and correlations: what's out there? Error or diagnostic variation? *Arch Pathol Lab Med* 2006;130(5):620-5.
17. Gostel R. HyperCard to SPSS: improving data integrity. *Comput Nurs* 1993;11(1):25-8.

18. Tran P, Morrison SG, Lade JA, Haw CS. An integrated approach to surgical audit. *ANZ J Surg* 2011;81(5):313-4.

19. Featherstone I, Keen J. Do integrated record systems lead to integrated services? An observational study of a multi-professional system in a diabetes service. *Int J Med Inform* 2012;81(1):45-52.

20. Veen EJ, Janssen-Heijnen ML, Leenen LP, Roukema JA. The registration of complications in surgery: a learning curve. *World J Surg* 2005;29(3):402-9.

21. Pearce C, Haikerwal MC. E-health in Australia: time to plunge into the 21st century. *Med J Aust* 2010;193(7):397-8.



TissuPath
Histopathology & Cytogenetic Services

T-44646
Surname, Firstname

DOB: **SEX**: M

Collection: 10/11/2008

TEST : T021 - SYNOPTIC (3/3)

SYNOPTIC REPORT: CARCINOMA OF PROSTATE

Specimen type: Robotic radical prostatectomy -- tissue banking correctly identified.

Tumour type: Index tumour:transition zone
Other tumour foci: not identified

Gleason score: Index tumour: 3+4(5%)=7
Other tumour foci: not identified

Site: See volumetric analysis.

Unifocal/Multifocal: unifocal

Perineural invasion: not identified

Lymphovascular invasion: not identified

Extraprostatic extension: not identified

Seminal vesicle involvement: not identified

Pathological resection margin status Clear

Presence of PIN: minimal

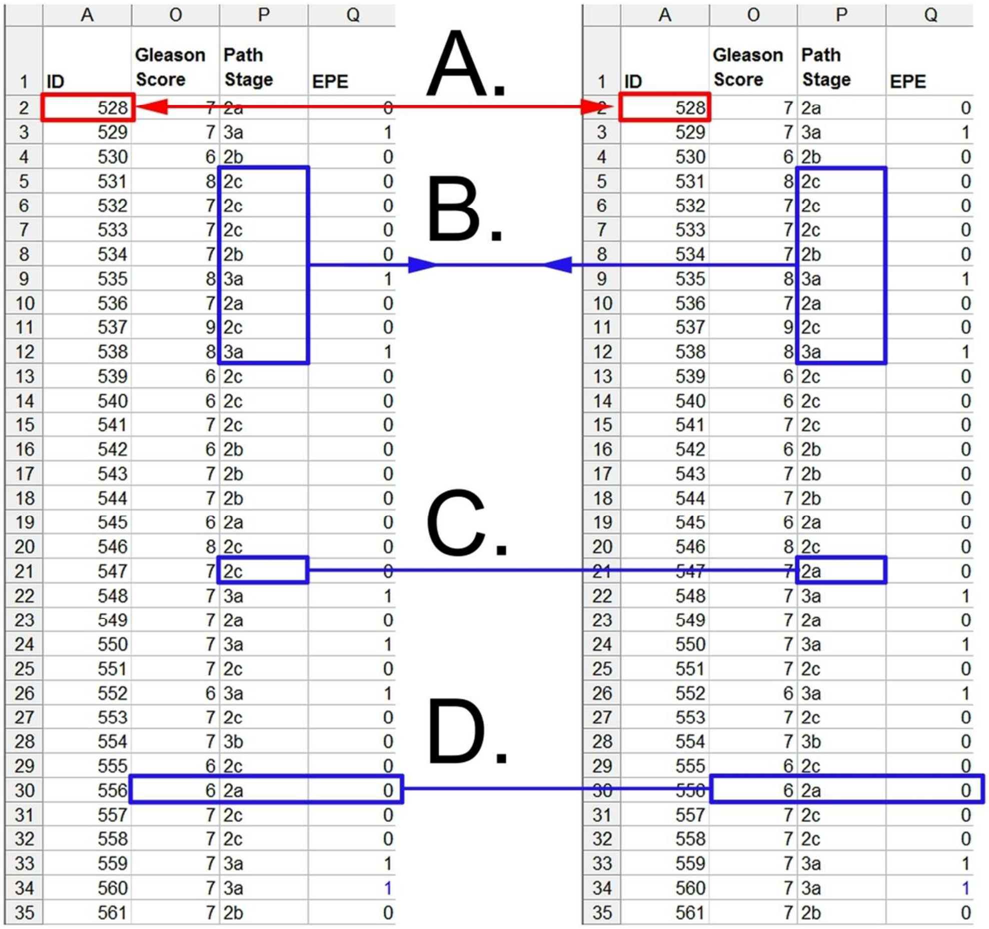
Lymph nodes: not taken

Tumour Volume: Transition Zone (green) 0.23 cc

Pathologic stage: pT2a.

	A	B	C	D	E	F	O	P	Q	V	W	X	Y	Z	AA
1	ID	Last Name	First Name	Surgeon	DOB	D of Surgery	Gleason Score	Path Stage	EPE	Tumour Vol	PN invasion	LV invasion	PIN	Multifocal	Margins
2	528	Last Name	First Name	###	######	#########	7	2a	0	4.95	1	0	1	1	0
3	529	Last Name	First Name	###	######	#########	7	3a	1	5.14	0	0	0	0	0
4	530	Last Name	First Name	###	######	#########	6	2b	0	6.24	1	0	1	1	0
5	531	Last Name	First Name	###	######	#########	8	2c	0	1.95	1	0	1	1	0
6	532	Last Name	First Name	###	######	#########	7	2c	0	0.89	1	0	1	1	0
7	533	Last Name	First Name	###	######	#########	7	2a	0	0.23	0	0	1	0	0
8	534	Last Name	First Name	###	######	#########	7	2b	0	2.64	1	0	1	1	0
9	535	Last Name	First Name	###	######	#########	6	3a	0	0.55	1	0	1	0	0
10	536	Last Name	First Name	###	######	#########	7	2a	0	10.17	1	1	1	1	0
11	537	Last Name	First Name	###	######	#########	9	2c	0	0.74	0	0	1	1	0
12	538	Last Name	First Name	###	######	#########	8	3a	1	4.35	1	0	1	0	1
13	539	Last Name	First Name	###	######	#########	6	2c	0	0.55	1	0	1	1	0
14	540	Last Name	First Name	###	######	#########	6	2c	0	0.70	1	0	0	1	0
15	541	Last Name	First Name	###	######	#########	7	2c	0	5.06	1	0	1	0	0
16	542	Last Name	First Name	###	######	#########	6	2b	0	2.68	0	0	1	1	0
17	543	Last Name	First Name	###	######	#########	7	2b	0	0.00	1	0	1	1	0
18	544	Last Name	First Name	###	######	#########	7	2b	0	0.94	1	0	1	1	0
19	545	Last Name	First Name	###	######	#########	6	2a	0	0.99	1	0	1	1	0
20	546	Last Name	First Name	###	######	#########	8	2c	0	5.93	1	0	0	1	1
21	547	Last Name	First Name	###	######	#########	7	2c	0	4.45	1	0	1	1	0
22	548	Last Name	First Name	###	######	#########	7	3a	1	4.95	1	0	1	0	0
23	549	Last Name	First Name	###	######	#########	7	2a	0	5.14	1	0	1	0	1
24	550	Last Name	First Name	###	######	#########	7	3a	1	6.24	0	0	1	1	1
25	551	Last Name	First Name	###	######	#########	7	2c	0	1.95	1	1	1	0	0
26	552	Last Name	First Name	###	######	#########	6	3a	1	0.89	0	0	1	0	1
27	553	Last Name	First Name	###	######	#########	7	2c	0	0.23	1	0	1	1	1
28	554	Last Name	First Name	###	######	#########	7	3b	0	2.64	1	0	1	1	0
29	555	Last Name	First Name	###	######	#########	6	2c	0	0.55	1	0	0	1	0
30	556	Last Name	First Name	###	######	#########	6	2a	0	10.17	1	0	1	0	0
31	557	Last Name	First Name	###	######	#########	7	2c	0	0.74	0	0	1	1	0
32	558	Last Name	First Name	###	######	#########	7	2c	0	4.35	1	0	1	1	0
33	559	Last Name	First Name	###	######	#########	7	3a	1	0.55	1	0	1	1	1
34	560	Last Name	First Name	###	######	#########	7	3a	1	0.70	1	0	1	1	1
35	561	Last Name	First Name	###	######	#########	7	2b	0	5.06	1	0	0	0	0

Fig. 1 Schematic representation of the digital import of pathology data. Structured ‘synoptic’ reports facilitated digital recognition of relevant pathology fields. (A) Demographics data was used to link reports to individual patients in the database and (B) individual data were then extracted from the report and directed to populate relevant fields in the main database.
156x90mm (300 x 300 DPI)



	A	O	P	Q
1	ID	Gleason Score	Path Stage	EPE
2	528	7	2a	0
3	529	7	3a	1
4	530	6	2b	0
5	531	8	2c	0
6	532	7	2c	0
7	533	7	2c	0
8	534	7	2b	0
9	535	8	3a	1
10	536	7	2a	0
11	537	9	2c	0
12	538	8	3a	1
13	539	6	2c	0
14	540	6	2c	0
15	541	7	2c	0
16	542	6	2b	0
17	543	7	2b	0
18	544	7	2b	0
19	545	6	2a	0
20	546	8	2c	0
21	547	7	2c	0
22	548	7	3a	1
23	549	7	2a	0
24	550	7	3a	1
25	551	7	2c	0
26	552	6	3a	1
27	553	7	2c	0
28	554	7	3b	0
29	555	6	2c	0
30	556	6	2a	0
31	557	7	2c	0
32	558	7	2c	0
33	559	7	3a	1
34	560	7	3a	1
35	561	7	2b	0

	A	O	P	Q
1	ID	Gleason Score	Path Stage	EPE
2	528	7	2a	0
3	529	7	3a	1
4	530	6	2b	0
5	531	8	2c	0
6	532	7	2c	0
7	533	7	2c	0
8	534	7	2b	0
9	535	8	3a	1
10	536	7	2a	0
11	537	9	2c	0
12	538	8	3a	1
13	539	6	2c	0
14	540	6	2c	0
15	541	7	2c	0
16	542	6	2b	0
17	543	7	2b	0
18	544	7	2b	0
19	545	6	2a	0
20	546	8	2c	0
21	547	7	2a	0
22	548	7	3a	1
23	549	7	2a	0
24	550	7	3a	1
25	551	7	2c	0
26	552	6	3a	1
27	553	7	2c	0
28	554	7	3b	0
29	555	6	2c	0
30	556	6	2a	0
31	557	7	2c	0
32	558	7	2c	0
33	559	7	3a	1
34	560	7	3a	1
35	561	7	2b	0

Fig. 2 Schematic representation of the comparison of a dataset imported digitally and in parallel to a manually entered dataset. (A) Records were linked using unique patient identifiers, and (B) pathology fields were individually compared. Concordant data were flagged for merging in order to eliminate duplicate data (C) Mismatches were used to identify errors in the manual entry dataset. (D) We compared across all pathology fields for individual patients.
97x90mm (300 x 300 DPI)